复杂生物网络及
非编码 RNA 参与的双色网络

刘长宁 孙世伟 赵屹 卜东波

摘要： 非编码核糖核酸（noncoding RNA, ncRNA）和复杂生物网络是目前生物学研究，特别是基因组研究领域快速发展的两个方向，也是生物信息学研究所关注的热点。本文分别对非编码核糖核酸研究以及复杂生物网络研究的相关背景和最新研究进展，特别是生物信息学在其中的应用进行了介绍，并进一步讨论了将非编码核糖核酸引入复杂生物网络研究的重要意义以及下一步的可能工作方向。

关键词： 生物信息学；蛋白质相互作用网络；非编码 RNA 基因；复杂网络分析

1 引言

继 1995 年流感嗜血杆菌基因组全序列测序完成之后 [1]，一系列生物的全基因组被测定，特别是在 2004 年人类基因组序列测定的基本完成 [2] 标志着生物学研究后基因组时代到来。作为生物遗传信息的载体，基因组全序列的测定完成给我们提供了丰富的信息，但同时也给我们带来了新的问题。传统观念认为生物的复杂性是由编码蛋白质基因的规模决定的，但通过物种间基因组序列的比对分析，我们发现人类和线虫虽然在生命的形态和复杂程度上有着巨大的差别，但两者编码蛋白质基因的数目相差却并不大 [3]。那么究竟还有什么原因造成了人类和线虫之间这么巨大差别呢？已有研究表明生命体可以表示为一个复杂的动态变化的网络。网络中的节点是各种生物分子，如脱氧核糖核酸（DNA）、核糖核酸（RNA）、蛋白质，它们之间的关联表示为边（无向边，如蛋白相互作用）或箭头（有向边，如基因调控关系）[4,5]。系统生物学认为这种网络关系正是生命复杂性的源泉，各种复杂的生命现象都是由网络中各节点之间的不同相互作用、调控关系的组合和动态变化产生的。生命复杂度的增加可以通过两种不同的途径：首先是增加网络中的节点数目，即在网络中添加更多的蛋白质和核糖核酸从而扩大网络的规模；更重要的是通过加强网络中节点间相互作用动态调整的能力和引入新的节点间相互作用机制。因此，对于生命的研究需要从整体出发，研究生命网络中的这些复杂的相互作用和调控关系，而不是仅仅孤立地研究一个个网络中的节点 [6]。

非编码核糖核酸基因是指转录后无需翻译成蛋白质，而直接以核糖核酸形式行使生物功能的基因 [7,8]。非编码核糖核酸在包括染色体表观遗传修饰、信使核糖核酸（mRNA）转录和降解、蛋白质的运输、核糖核酸的加工修饰等多个重要环节发挥功能 [9-12]，同多种疾病以及肿瘤的发生密切相关 [13-16]。随着大量非编码核糖核酸在多种模式生物中的相继发现，对非编码核糖核酸的研究已经成为生物学研究中的重点和热点 [17-19]。生物网络中非编码核糖核酸的加入大大增加了生命的复杂性：首先它在网络中增加了数以万计的节点，扩大了网络的规模；更重要的是它在网络中增加了各种新的相互作用机制，例如近年来才发现的微型核糖核酸（miRNA）[1]对信使核糖核酸的互补抑制作用就在生物复杂网络中引入了一个全新的转录后调控层次 [10]。由于认识到非编码核糖核酸的出现对生物网络可能的重大影响，对非编码基因的研究现在已经从寻找新的非编码基因，研究探索非编码基因功能向研究和建立

[1]生物体内源长度约为 20－23 个核苷酸的非编码小核糖核酸。

非编码基因和编码基因的混合网络这个方向扩展。虽然混合网络的研究还刚刚开始，但必将成为非编码基因研究的新热点。在本文中，我们将结合近几年来我们在非编码核糖核酸以及生物复杂网络方面所做的工作，分别对非编码核糖核酸功能研究，复杂生物网络分析以及非编码核糖核酸参与的生物网络的构建几个方面的已有工作以及下一步的可能研究方向进行介绍。

2　非编码核糖核酸与核糖核酸组学

在高等生物和人的基因组中非编码区占到基因组序列的大部分，如人类基因组和小鼠基因组中的编码蛋白质的序列只占约 3-5%，其余约 95-97%为非编码区 [20,21]。这些区域一度被认为是没有任何功能的“垃圾 DNA”。但从生物进化的观点来看，非编码区序列随着生物体功能的完善和复杂化而明显增加的趋势表明，非编码区序列必定具有重要的生物功能。最近几年国内外学者对大规模转录组的相关研究日益深入。大量的实验数据表明基因组非编码区不但作为结合位点参与转录调控，而且还能转录出数目众多的非编码核糖核酸产物。相关研究包括：（1）大规模互补脱氧核糖核酸（cDNA）注释研究，如 2003 年，RIKEN 国际联盟在克隆分析小鼠全长互补脱氧核糖核酸时发现其中有近 4280 个全长互补脱氧核糖核酸是缺乏蛋白质编码读框的非编码核糖核酸基因 [17,22]；（2）基因芯片研究，如 2005 年 Affymetrix 公司在运用高密度的寡核苷酸芯片对 10 条人类染色体的转录组研究中证实了大量的非编码核糖核酸基因的存在[23]；（3）实验核糖核酸组学，如 2006 年中科院生物物理所陈润生实验室在对线虫的微型核糖核酸研究中发现了大量新的非编码核糖核酸，包括两类新的非编码核糖核酸、小核样核糖核酸（small nuclear-like RNA，snlRNA）和柄部突出核糖核酸（stem-bulge RNA，　sbRNA）[24]。当然还有大量的其他类似工作不能一一列举。到目前为止各国科学家已经在包括小鼠、果蝇、拟南芥、水稻、古细菌甚至大肠杆菌等多种生物中发现了大量的非编码核糖核酸 [18,19,25-28]。

已有研究发现这些长短不一，结构各异的非编码核糖核酸在生物体中发挥着各种不同功能，如小核核糖核酸（small nuclear RNA, snRNA）参与信使核糖核酸剪接 [29]；小核仁核糖核酸（Small nucleolar RNAs，　snoRNA）参与核糖体核糖核酸（ribosomal RNA，rRNA）的甲基化和假尿嘧啶化加工 [12]；向导核糖核酸（guide RNA，gRNA）参与核糖核酸编辑[30]；信号识别颗粒核糖核酸（The Signal Recognition Particle RNA，SRP-RNA）参与蛋白质的细胞定位[11]；端粒核糖核酸参与脱氧核糖核酸端粒合成并影响细胞的寿命[31]；转移信使核糖核酸（transfer-messenger RNA, tmRNA）参与终止受损的信使核糖核酸的蛋白质合成过程[32]；Xist 能使 X 染色体失活 [33]；piRNA[2]参与调控染色体表观遗传修饰等[9]。另外，在最近对多种疾病和肿瘤的医学研究中也发现了大量肿瘤和疾病特异表达的非编码基因，如在非小细胞肺癌中高表达的非编码核糖核酸基因 MALAT-1[16]，在前列腺癌中异常表达的非编码核糖核酸基因 PCGEM1 等[15]。相对于已知功能的非编码核糖核酸，我们对于绝大部分非编码核糖核酸的功能可以说近乎一无所知，如何研究这些非编码核糖核酸的调控与功能已经成为生物学研究的新挑战。中外科学家都已经注意到了以此为研究对象的核糖核酸组问题，早在 1998 年我国科学家金由辛就在第 109 次香山科学会议上提出了“功能核糖核酸组研究计划”，国外在 2000 年左右也已经开始了大规模的实验和计算核糖核酸组学研究，在 2001 年～2006 年，这个领域的重要发现 5 次被《科学（*Science*）》归入当年的年度十大科学发现。以非编码核糖核酸为研究主题的核糖核酸组学研究已经成为实验生物学和生物信息学领域的热点。

[2] Piwi-interacting RNA，一类小型核糖核酸分子，长度大约是 29 到 30 个核苷酸。只表现在哺乳动物的睾丸中，并且可以和 Piwi 蛋白结合形成 piRNA 复合物（piRNA complexes，piRCs）。与核糖核酸沉默（RNA silencing）作用有关

1.1 非编码 RNA 基因数据库 NONCODE 的建立

随着对非编码基因的日益重视和相关研究的深入开展，越来越多的非编码基因新成员和非编码基因新类被发现，收集、组织非编码基因相关信息的数据库也开始出现。这些数据库

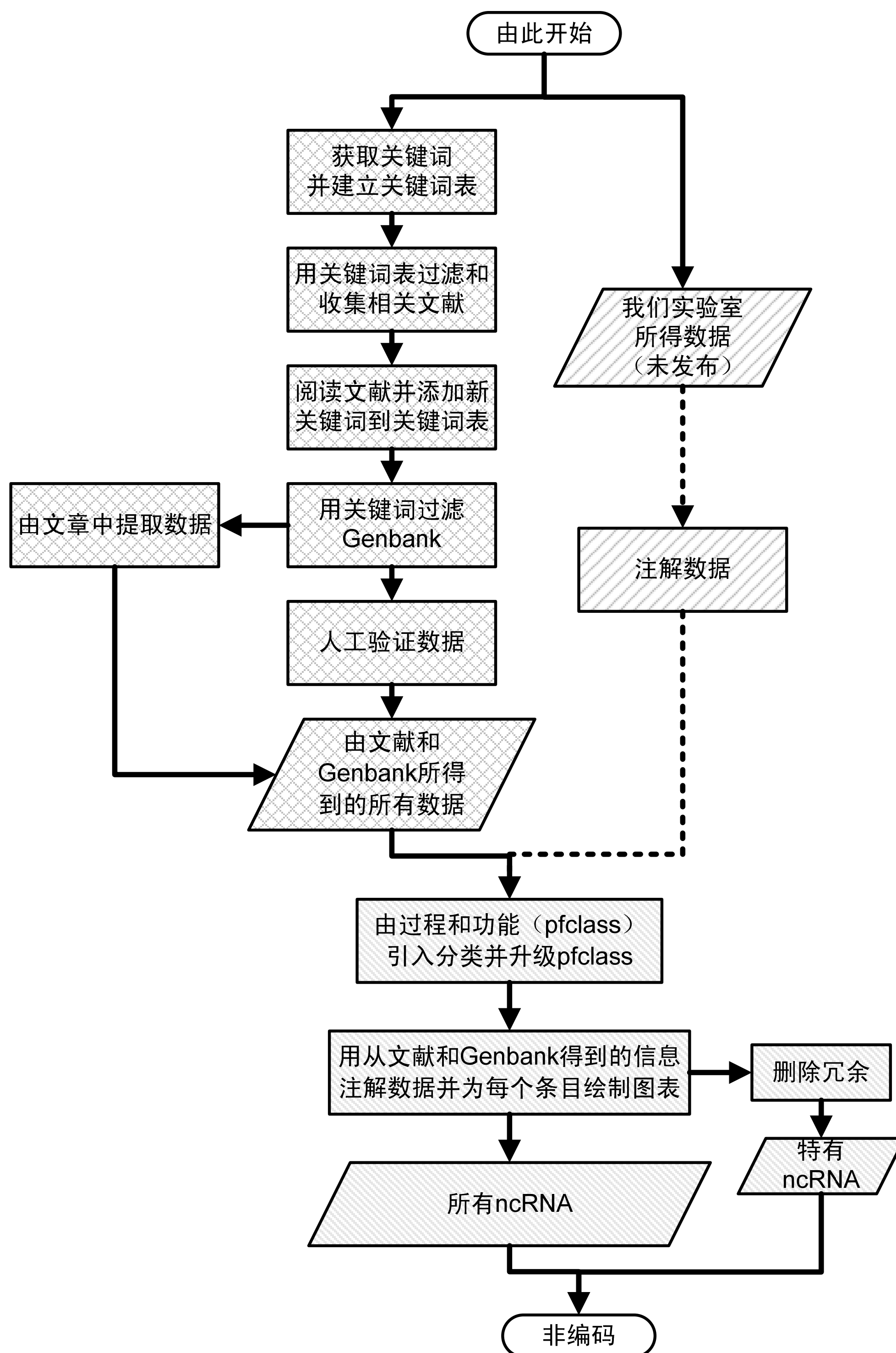


图1. NONCODE 数据收集处理流程

中有的只关注于某一类非编码基因，如 SRP RNA，tmRNA，和 RNase P RNA[3]，有的则是收

[3] 核糖核酸酶 P 中的核糖核酸组分,普遍存在于古生菌、细菌、真核 及叶绿体、线粒体中的一种核糖核酸内切酶.

集了各种非编码基因数据，如“Small RNA Database”、“NoncodingRNA Database”以及“Rfam Database”[34-39]。然而这些数据库都存在着一些问题。首先是由于它们收集的数据往往是通过人工从文献中获取，所以收集的非编码基因数据不论是从数量还是种类上来说都有很多遗漏。另一个更严重的问题是他们都没有一个统一的对非编码基因分类注释的系统，而这个问题带来的麻烦更加危险。NONCODE 就是在这样的背景下开始建设的。 一方面，NONCODE 采取了计算机自动过滤 GenBank[40]数据然后人工检查确认的工作方式。这样既提高了收集数据的全面性和准确性，又保证了工作效率。另一方面，为了解决非编码基因缺乏统一分类体系的问题，我们提出了一套以非编码基因所参与的细胞生化过程和在此过程中发挥的功能为标准的全新的、统一的分类体系——“过程功能”分类系统。在第一版 NONCODE 数据库中我们共收集了除转移核糖核酸（transfer RNA，tRNA），核糖体核糖核酸以外所有种类的非编码基因数据共计 5339 条非冗余记录，涉及 861 个物种，遍及真细菌、古细菌和真核生物界[41]。

为了高效而且全面地收集非编码基因数据，我们以 PubMed 为起点设计了一套计算机自动分析辅助人工确认的数据收集流程（见图 1）。PubMed 是美国国家医学图书馆所属的国家生物技术信息中心开发的互联网生物医学信息检索系统，覆盖了全世界 70 多个国家 4300 多种主要生物医学期刊的摘要和部分全文。我们用关键字表检索 PubMed，检索得到的文献通过手工检查，以确认文献和非编码基因相关。通过阅读这些非编码基因相关的文献，进一步得到新的非编码基因关键字。我们根据这些新的关键字更新关键字表，然后用新的关键字表自动过滤 GenBank 中的 GB 格式文件。GenBank 由美国国立生物技术信息中心建立和维护，其中包含了所有已知的核酸序列和蛋白质序列以及与它们相关的文献著作和生物学注释。每个 GB 格式文件包含了对序列的简要描述、科学命名、物种分类名称、参考文献、序列特征表以及序列本身。序列特征表里包含对序列生物学特征注释如：编码区、转录单元、重复区域、突变位点或修饰位点等。根据 GB 文件中的这些注释和我们的非编码基因相关关键字表我们可以粗筛出可能的非编码基因，并能对筛选出的候选非编码基因进行初步分类。所有 GB 文件被分为细菌类、病毒类、灵长类、啮齿类以及 EST 数据、基因组测序数据、大规模基因组序列数据等 16 类。我们的搜索主要针对其核酸库中的真核、原核、细菌、病毒、类病毒等几类。搜索得到的数据被导入 MySQL 数据库中等待人工检查确认。经过人工确认其为真实的非编码基因数据则对其进行一系列注释工作。同样，整个注释过程基本由计算机自动完成，少数特殊情况计算机将提示需要人工确认。最后，我们在这个数据库的基础上建立了一个界面友好、功能全面的网络接口（www.noncode.org），提供数据浏览、关键字搜索、序列在线 Blast 查询、数据下载等一系列服务。

在现有的非编码基因的命名中，有的非编码基因是根据其在细胞中的定位来命名的，如小核核糖核酸（在细胞核中），小核仁核糖核酸（在核仁中）[29,42]；有的非编码基因是根据功能来命名的，如 pRNA（package RNA，组装核糖核酸），向导核糖核酸[43,44]；更有甚者，直接用非编码基因的沉降系数来命名，如 6S RNA，5.3S RNA 等[45]。这些不同的命名方法导致同一类非编码基因由于来自不同的实验室往往会有多个名字，还有很多名字相同但功能完全不相关的非编码基因出现。我们根据非编码基因参与的细胞生化过程及其发挥的功能制定了一套统一的分类系统，希望通过这种分类避免以前发生的混乱现象，同时便于研究者从分类直接了解某一类非编码基因的功能。在 NONCODE 数据库的"过程功能"分类系统中，细胞过程指以脱氧核糖核酸、核糖核酸、蛋白质三者为作用底物的生物反应，如脱氧核糖核酸的复制、修饰，核糖核酸的可变剪接、甲基化修饰，蛋白质的输运、降解等，每条非编码基因都以其在一个细胞过程中所行使的功能来命名，整个命名由下划线所连接的两级到三级关键字给定。第一级关键字是 DNA，RNA，Protein，代表在一个细胞过程中哪个分子类型

为关键成分，第二个关键字描述了具体的一个过程，如果这个过程存在更多的细节分支，则用第三个关键字来进一步解释具体的功能。例如，非编码基因 snRNA U1，它参与了信使核糖核酸剪接的过程，主要分子是 RNA，过程是对 RNA 的加工处理，更细节的具体过程是 splicing（剪接），因此 snRNA U1 将会被分配到 RNA_processing_splicing（核糖核酸-过程-剪接）这个类里，而 RNase P RNA 参与了转移核糖核酸 5’端成熟的过程，切割转移核糖核酸前体 5'端，因此分配到 RNA_processing_ cleavage （核糖核酸-过程-解理）。“过程功能”分类系统是第一个尝试把非编码基因参与的过程及行使的功能整合在一起的一个分类系统。将来随着我们对非编码基因认识的深入，NONCODE 数据库的内容也会进一步地扩充，这个分类系统也会得到进一步的完善，使得数据库能充分地得到利用。有关"过程功能"分类系统的详细情况参见表 1。

表1. ″过程功能″分类系统

Pf（过程功能）类	相对应的传统类
DNA_ imprinting	XIST, roX , H19, MHM, KvLQTI-AS, Tsix, Air
DNA_ packaging	pRNA
DNA_ repair	RNA a, b, c, d
DNA_ replication _initiation	RNAII
DNA_ replication _regulation	ctRNA, RNA I
DNA_ replication _ repression	incA, RNA I
DNA_ stability	telomerase RNS
DNA_ transcription _initiation	RNA II
DNA_ transcription _regulation	Inc RNA, copA RNA, SRA
DNA_ transcription _ regulation of RNA polymerase	6S RNA, 7SK
DNA_ transcription _repression	RNAI, GevB RNA
RNA_ editing	gRNA
RNA_ modification _ methylation	snoRNA
RNA_ modification _ methylation & pseudouridylation	scaRNA
RNA_ modification _ pseudouridylation	snoRNA
RNA_ processing _ cleavage	RNase P RNA, RNase MRP RNA, snoRNA
RNA_ processing _ splicing	snRNA, self-splicing ribozyme RNA, PAN
RNA_ reverse _ transcription	msrRNA
RNA_ translation _ enhancement	csrB RNA, DsrA RNA
RNA_ translation _ regulation	ANTI-RAF1, RprA, sok RNA, VA RNA, RyhB, sar RNA, NaPi-2b1, 5.3S RNA, aHIF
RNA_ translation _ suppression	miRNA, DicF, Spot 42, Finp, MicF, OxyS, flmB, PrrB_RsmZ, NTT, GevB DNA, etc.
RNA_ translation _ surveillance	tmRNA
RNA _translocation	ScYC RNA, hsr-omega RNA, Xlsirt
Protein _ transportation	SPR_7SL, RNA, SRP_4.5S RNA
Miscfunction _ mRNAlike	BORG, IGF2AS, CR20, meuRNA, Rian, Ks-1, GNAS1-as RNA, IPW, etc.
Miscfunction _ snm	Bsr RNA, Y RNA, dsrB, vault RNA, 4.5S RNA, 6Sa RNA, G8, etc.

1.2 微型核糖核酸编码（miRNA-encoding）非编码基因的预测和验证

最近几年的几个重要模式生物的全基因组芯片实验和全长互补脱氧核糖核酸（cDNA）文库建设都发现基因组上存在着大量长的非编码转录本。它们和编码蛋白的信使核糖核酸有一些相似之处：长度都很长，都由 RNA 聚合酶 II 转录，转录后都存在剪接、加帽加尾现象，但是又没有蛋白编码框，因此被称为“信使核糖核酸样（mRNA-like）非编码基因”[46-51]。已经发现的信使核糖核酸样非编码基因的数目惊人，比如在 FANTOM 小鼠全长互补脱氧核糖核酸文库中发现有约 4000 个全长互补脱氧核糖核酸是缺乏蛋白质编码读框的信使核糖核酸样非编码基因[52]；类似的，在人类全长互补脱氧核糖核酸文库中发现了近 5800 个信使核糖核酸样非编码基因[53]。少数信使核糖核酸样非编码基因的功能已经得到证实，例如，马拉伦斯（Marahrens）等人发现雌性小鼠上的信使核糖核酸样非编码基因 Xist 如果被敲除，

将会影响小鼠 X 染色体的选择性失活[54]；杨（Young）等人发现对新生小鼠视网膜细胞中的信使核糖核酸样非编码基因 TUG1 的微型核糖核酸干涉会导致小鼠眼发育的畸形[55]；威廉汉姆（Willingham）等人发现小鼠信使核糖核酸样非编码基因 NRON 是转录因子 NFAT 的抑制子[56]。然而绝大部分信使核糖核酸样非编码基因的功能和作用机制仍然是未知的。微型核糖核酸（MicroRNA）是一种广泛存在于高等动物和植物中的微小的非编码基因，通过控制信使核糖核酸的稳定性或抑制信使核糖核酸的翻译对生命活动起到重要调控作用[57,58]。根据基因组的定位，微型核糖核酸可以被分为三类：（1）位于蛋白质转录区内含子的微型核糖核酸；（2）位于非编码转录区内含子的微型核糖核酸；（3）位于非编码转录区外显子的微型核糖核酸。我们推测存在更多的信使核糖核酸样非编码核糖核酸，在它们的外显子中编码微型核糖核酸。它们构成了一类特殊的非编码核糖核酸，我们称之为“微型核糖核酸编码的非编码核糖核酸”（microRNA-encoding ncRNA（me-ncRNA））。在本文中，我们通过分析老鼠基因组中一些编码已知微型核糖核酸的 20 条 me-ncRNA，设计了一种新的预测方法（PriMir），并利用该方法在 FANTOM3 数据库的 34030 条微型核糖核酸样非编码核糖核酸（microRNA-like ncRNA ）中预测了 65 条新的候选 me-ncRNA，其中 24 条得到了生物实验的证实。我们对这些已知的 me-ncRNA 和所预测的候选 me-ncRNA 进行了进一步分析，发现它们都含有一些保守的模体（motif）。我们的这项工作发现的 me-ncRNA 是一类新的非编码核糖核酸。我们还对一些功能未知信使核糖核酸样非编码核糖核酸给出了新的解释。

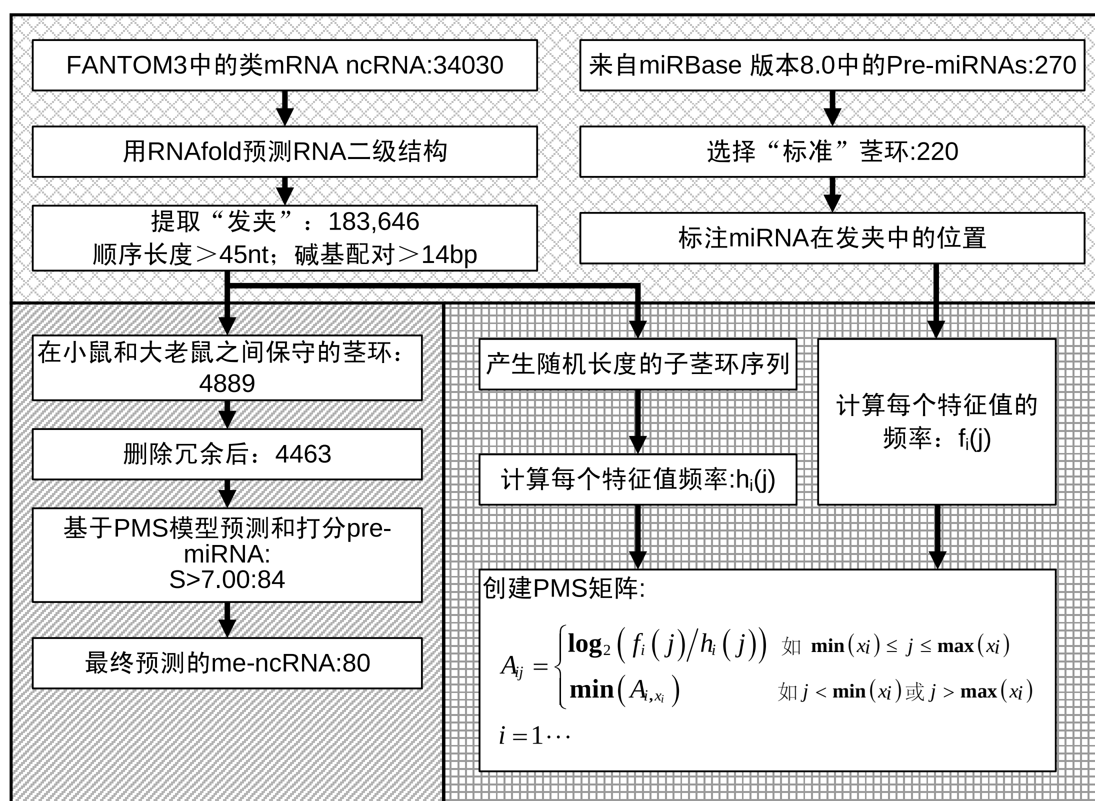


图2. PriMir 的流程图

我们还用 PriMir 来从所有的信使核糖核酸样非编码核糖核酸中寻找 me-ncRNA。PriMir 通过扫描所有信使核糖核酸样非编码核糖核酸的二级结构过滤出序列长度、碱基配对数目和已知微型核糖核酸前体（pre-miRNA）符合的茎环结构；然后通过小鼠大鼠间保守性分析过滤出所有在小鼠和大鼠之间保守的茎环结构；最后通过 PriMir 打分矩阵（PriMir score matrix，PMS matrix）预测出所有可能的微核糖核酸前体基因以及它们对应的 me-ncRNA。图 2 显示的是一个 PriMir 方法的流程图。为了建立训练集，我们分析了从 miRBase8.0 中得到的 270

条微核糖核酸前体，滤掉了其中茎环结构长度太短（小于 45nt[4]）以及微型核糖核酸成熟体序列位置特殊（在茎环结构的环上）的特例，这样得到了我们的训练集共 220 条已知的微核糖核酸前体序列。我们还需要建立一个由非微核糖核酸前体茎环结构组成的背景集。我们用 RNAfold 预测了 FANTOM3 中所有 34030 条信使核糖核酸样非编码核糖核酸的二级结构，然后 PriMir 根据两个条件从这些二级结构上提取满足下述两个条件的茎环结构：（1）茎环结构的序列长度大于 45nt；（2）茎环结构上的配对碱基数大于 28。这样我们得到了 184000 个茎环结构。这 184000 个茎环结构中存在着我们需要鉴定发现的真实的未知微核糖核酸前体，但大部分肯定都是非微核糖核酸前体茎环结构，因此我们用这 184000 个茎环结构作为背景集。在确定了训练集和背景集之后我们就可以通过分析在训练集和背景集中 11 个特征参数取值的差异建立 PMS 矩阵（见图 2）。对从所有 34030 条信使核糖核酸样非编码核糖核酸的二级结构上提取的 184000 个茎环结构我们根据其序列在小鼠和大鼠之间的保守性进行了进一步过滤。为了确定保守性的阈值，我们将训练集里 220 条已知的小鼠微核糖核酸前体与大鼠的基因组用 BLASTN 进行了比对。结果显示其中 160 条微核糖核酸前体满足以下两个标准：（1）比对上的序列长度超过 50nt；（2）比对的辨识（identity）值大于等于 98%。因此 PriMir 根据这两个标准对 184000 个茎环结构进行进一步过滤，得到了 4463 条在小鼠大鼠间保守的茎环。其中包括 18 条已知的微核糖核酸前体。然后 PriMir 用 PMS 矩阵对 4463 条保守茎环打分。为了减少假阳性的数量，PriMir 打分“7”被用作“截断（cutoff）”值。这是一种严格的评判标准，因为训练集里 220 条已知的小鼠微核糖核酸前体仅有 73%的评分在这个数值以上。这些 PriMir 打分分值大于等于 7 的微核糖核酸前体被认定为可能的微核糖核酸前体候选者。这样我们从 4463 条保守茎环中最终得到 84 条微核糖核酸前体候选基因，它们对应着 80 条可能的微核糖核酸前体。其中 15 条微核糖核酸前体属于 miRBase8.0 收录的已知微核糖核酸前体，它们对应着 15 条已知 me-ncRNA。因此我们称剩下的 69 条茎环及其对应的 65 条信使核糖核酸样非编码核糖核酸为微核糖核酸前体和 me-ncRNA 候选基因。

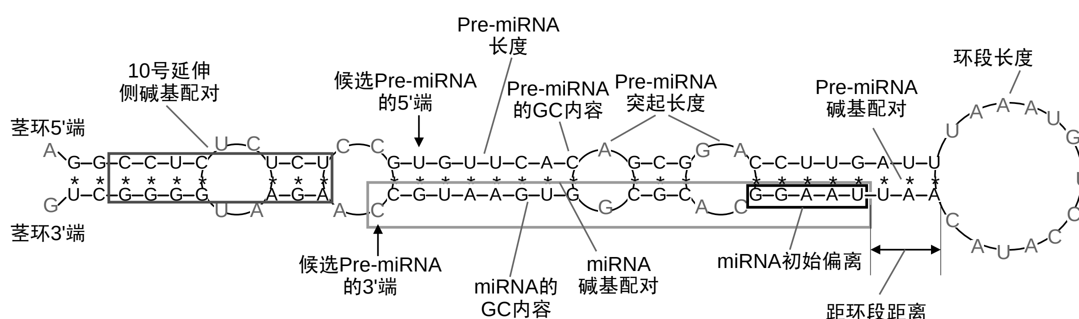


图3. PriMir 用的前体微核糖核酸的 11 个特征（用短线连接文字表示）示意图

PriMir 方法的灵敏性可以从对 20 条已知 me-ncRNA 的预测情况得到。由于我们预测出了其中的 15 条，灵敏性应该在 75%以上。由于信使核糖核酸样非编码核糖核酸的表达往往是组织特异性的或发育阶段特异性的，而且信使核糖核酸样非编码核糖核酸的表达水平往往很低，对于预测结果特异性的估计就比较困难。为了估计 PriMir 预测结果的特异性，我们设计了一张有 168 个 26-nt 探针的微阵列（microarray）。这些探针对应着预测的 84 条微核糖核酸前体的茎对应的双臂。为了防止杂交时长的核糖核酸的信号干扰，我们从提取的总核糖核酸中滤掉了长度大于 200nt 的核糖核酸分子。对初生小鼠脑组织和胸腺组织、2 个月雄性成年鼠脑组织以及 15 天小鼠胚胎提取核糖核酸杂交微阵列信号，结果分析显示有 46

[4] 核苷酸

个探针有显著信号。它们对应着 46 条不同的微型核糖核酸，40 条不同的微核糖核酸前体和 39 条 me-ncRNA（其中包括 15 条已知 me-ncRNA），其中有 6 条微核糖核酸前体的双臂都能检测到显著信号。15 条属于 miRBase8.0 收录的已知微型核糖核酸中有 14 条能够检测到显著信号，说明我们设计的微阵列运行良好。对于经过微阵列检测出来的 32 条新的微型核糖核酸，我们通过检索 miRBase9.0 发现其中 5 条已经被最新收录，剩下的 27 条微型核糖核酸我们对它们采用 Stem-loopRT-PCR 加测序的方法进行了进一步严格的检验，结果显示所有新的微型核糖核酸都是真实的。这样 65 条 me-ncRNA 候选基因中有 24 条通过了我们严格的实验检验。在我们的工作正在进行时，又有 10 条微型核糖核酸被其他实验室发现。这些微型核糖核酸对应着我们工作中的 5 条 me-ncRNA 候选基因和 4 条已知 me-ncRNA 基因。其中一条 me-ncRNA 候选基因通过了我们的 PriMir 预测，但我们的微阵列和 RT-PCR 加测序的方法没有检测出来。这样如果我们把我们自己的微阵列和 RT-PCR 加测序实验验证以及其他实验室发表的文献支持都算在内的话，PriMir 方法的特异性应该是 50％（(39+1)/80）。当然如果在我们的实验中考察更多的小鼠组织和发育时期样本，我们相信将会得到更高的特异性分值。

进一步，我们对 me-ncRNA 的序列保守性和序列 motif 进行了分析。为了衡量 me-ncRNA 的保守性，我们根据小鼠基因组相对于 17 种脊椎动物的 PhastCons 打分对 me-ncRNA 的每一个碱基的保守性评分，然后用整条 me-ncRNA 所有碱基评分的均值（average PhastCons scores，APCSs）做为一条 me-ncRNA 的保守性评分。65 条候选 me-ncRNAs 的 APCSs 平均为 41％，这个值远远高于 20 条已知的 me-ncRNA 的 APCs 均值（26％）。造成这 20 条已知 me-ncRNA 保守性低的原因可能是由于统计上的涨落，因为 20 条序列太少。另一个可能的原因是 me-ncRNA 对整条序列的保守性没有要求。因此我们又统计了 me-ncRNA 的微核糖核酸前体部分的序列保守性。结果显示已知和预测 me-ncRNA 的微核糖核酸前体部分的平均 APCs 分别是 88%和 72%，都大大高于整条 me-ncRNA 序列的保守性得分。通过分析已知和预测的 me-ncRNA 我们在 me-ncRNA 的内部找到了一个内部 motif IM1，其保守序列是 CNCTUNCTU（见图 4(a)）。我们根据 IM1 建立了一个位置权重矩阵（Positional Weight Matrix，PWM），然后用这个矩阵搜索了所有信使核糖核酸样非编码基因序列。矩阵打分阈值被定为保证 50%的确证 me-ncRNA（20 条已知的加上 24 条实验确证的）上有 IM1 出现。结果显示在 65%的已知 me-ncRNA 和 42%的预测 me-ncRNA 上有 IM1，而在所有的信使核糖核酸样非编码基因上有 23%的序列有 IM1（见图 4(b)）。因为内部 motif IM1 在信使核糖核酸样非编码基因上出现的比例也比较高，因此我们进一步分析了 IM1 在序列上出现的频率和 PriMir 方法对序列的打分之间的关系。我们用 PriMir 对一条信使核糖核酸样非编码核糖核酸上所有茎环结构打分的最高分作为此信使核糖核酸样非编码核糖核酸的得分。对所有信使核糖核酸样非编码核糖核酸的分析结果发现在序列上出现 IM1 的次数和序列的 PriMir 得分两者之间存在强相关性（决定系数（R-squared ）= 0.91，p 值= $2.2e^{-16}$）（见图 4 (b)），也就是说序列上存在微型核糖核酸的可能性（PriMir 得分）越高则序列上存在 IM1 的可能性和数目就越大。另一方面，茎环结构越保守则越可能是真实的微型核糖核酸。因此我们认为如果 IM1 确实和序列是否编码微型核糖核酸相关，那么加入保守性限制条件后，IM1 和 PriMir 得分间的相关性将会下降，这是因为在有了严格的保守性要求的前提下 PriMir 低分的茎环结构仍然可能编码微型核糖核酸，因而仍然存在大量的 IM1 motif。进一步分析的结果和我们预测一致，在具有保守茎环结构的 3670 条 mRNA-like ncRNA 的子集中，IM1 和 PriMir 得分间的相关性大大降低（决定系数（R-squared）= 0.1， p 值 = 0.03）（见图 4(b)），而其中出现 IM1motif 的序列的比例却比全部信使核糖核酸样非编码核糖核酸集合要高。综合以上分析我们认为 IM1 motif 确实和 me-ncRNA 编码微型核糖核酸存在关系，IM1 的发现将有助于我们对 me-ncRNA 的进一步预测，同时 IM1 在 me-ncRNA 编码微型核糖核酸过程中的

功能还需要我们进一步深入研究。

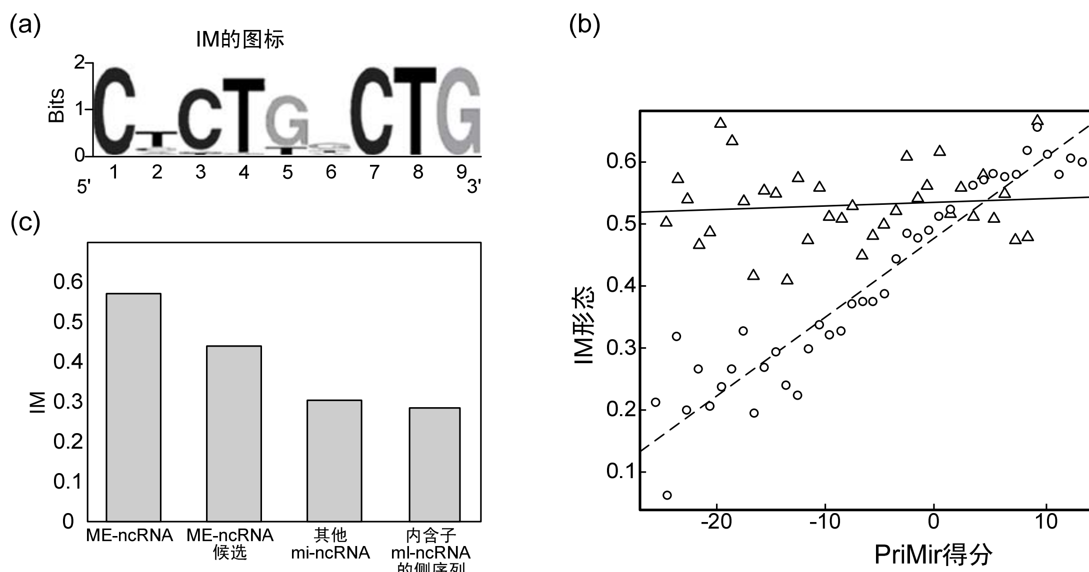


图4. Me-ncRNA 的内部 motif

3 复杂生物网络与系统生物学

系统生物学认为生命体是一个复杂的动态变化的网络，对于生命的研究需要从整体出发，研究生命网络中的这些复杂的相互作用和调控关系，而不是仅仅孤立地研究一个个网络中的节点[59]。最初的系统生物学研究由于实验技术限制主要是停留在对计算机模拟系统的理论研究。随着一系列生物的全基因组被测定，以及全基因组芯片、酵母双杂交、染色质免疫共沉淀微阵列等各种高通量实验技术的出现[60-62]，从全局观测整个生物网络的拓扑结构以及网络中节点的定量变化成为可能：通过分析全基因组序列，预测编码蛋白质基因和非编码基因，我们可以迅速确定生物网络中大部分的节点；通过酵母双杂交技术大规模检测蛋白质之间的相互作用，我们可以找到生物网络中的无向边；通过染色质免疫共沉淀微阵列技术检测转录因子在染色质上的特异结合位点，我们可以确定生物网络中的有向边；最后通过全基因组芯片等基因芯片技术我们可以动态地定量观测生物网络中各节点的表达水平。虽然这些高通量的实验技术不可避免地存在着各种噪音和测量偏差，它们仍然为系统生物学研究提供了基础。现在，系统生物学已经成为了一种新的生命科学研究的工作模式。它从多数据源整合出发，以网络分析为基础，通过统计学、信息学、人工智能等各种手段，对各种生命现象做出预测并指导传统生物实验对预测做出验证。这种新的工作模式极大地促进了生命科学研究的进展，使后基因组时代的系统生物学研究进入了一个高速发展的新时期。

人类基因组计划的完成是生命科学发展的一大步，下一步将由功能基因组学来研究已破译基因的功能并控制它们，最终为人类征服自然、战胜疾病服务。正如 Millenium Pharmaceutical 公司的罗伯特. 泰珀（Robert Tepper）所说，“我们知道了词典里面有什么，现在我们需要知道每个词的意思”。尽管基因序列的 99%已经被破译，但是只有 10%的基因的机能是已知的，如何获得更多基因功能成为功能基因组学的主要研究课题。很长时间以来，研究基因的功能都是针对单个基因来进行的，其思路是“序列→结构→功能”。认为一个基因表达一个蛋白质，一个蛋白质有一个结构，一个结构完成一个功能。相对于后基因组时代的功能基因组研究目标来说，这种“一次一个基因”的研究模式不但在效率上已经完全不能

适应要求，更严重的是这种研究方式本身就无法揭示生命活动的复杂性和本质。现在越来越多的研究表明，一个基因的单独表达往往不能主宰一个生物学事件的发生。生物的功能一般都是通过一批基因的同时表达，一批蛋白质的协同作用来实现的。在一个生物学事件中，存在着复杂的基因转录调控网络来控制相关基因的同时表达，还存在着各种蛋白质，甚至核糖核酸 互相结合的相互作用网络。所以改变原来的"一次一个基因"研究方式和"序列→结构→功能"思路，以系统生物学的观点，采用 "相互作用→网络→功能"新思路，整合基因和蛋白质的不同方面、不同层次的信息，进行基因功能分析，已经成为当前功能基因组研究的新方向。在对于编码基因的研究中，基于蛋白质相互作用网络以及基因转录调控网络的研究已经展现了网络研究的巨大威力：通过网络聚类寻找功能模块，根据网络邻居节点预测蛋白质功能，研究网络模体的拓扑结构和信号传导特性，这些基于网络的研究已经成为生物学研究的新武器。

3.1 蛋白质相互作用网络的谱分析方法

后基因组时代的一个巨大的挑战就是如何理解基因的信息是如何导致基因产物间相互协同作用，以及它们之间又是如何在时间和空间上行使生物功能，最终彼此间相互作用形成一个有机体。因此，发展一套可依赖的蛋白组学的方法来更好地理解蛋白功能是非常重要的。基因组学的方法已经被利用来根据序列特征推测大量的基因的功能。但是众所周知蛋白在生化层次上很少是单独起作用的，而是与其他的蛋白相互作用形成整体来实现细胞的某些特定的任务。系统的功能要比他们各部分分别体现的功能更丰富。传统上讲，蛋白相互作用的研究是对从遗传、生化和生理角度上讲在某个时刻的一些蛋白进行研究。现在我们认识到这种对细胞内的遗传和生化通路拼图式的研究已经阻碍了我们对细胞作为整体的生物过程的进一步的认识。而蛋白复合物、细胞通路、蛋白相互作用等基本的组成部分才对蛋白功能具有决定性的作用。所以，可以确信所有的生物过程从本质上更精确地说都是通过蛋白相互作用体现出来的。最近三年来发展出了高通量的相互作用的探测方法，比如酵母双杂交系统、基于质谱技术的蛋白纯化方法、具有相关信息的表达谱分析方法、遗传相互作用网络方法以及其他的基于基因相关性的计算模型的相互作用预测方法（基因融合和分裂、基因邻居和共出现基因等），它们在若干个生物（如酿酒酵母（S. cerevisiae）、秀丽线虫（Caenorhabditis elegans）和幽门螺旋杆菌（Helicobacter pylori））产生了数量可观的蛋白相互作用的大规模数据[63-75]。这些高通量大规模数据为更全面了解细胞中的遗传和生化现象开启了一扇大门。随后，几种方法被成功地应用于这方面的研究。比如，施维科夫斯基（Schwikowski）等和菱垣（Hishigaki）等成功地利用相互作用邻居来预测未知蛋白的功能。葛（Ge ，音译）等首次为具有相似表达谱的蛋白具有倾向于具有蛋白相互作用提供了证据。弗雷泽（Fraser）等揭示了具有保守的相互作用的蛋白与他们的突变率呈现负相关性[76-79]。所有的这些研究都预示着对于酿酒酵母的相互作用网络可能具有和其他复杂网络不同的性质。相互作用的拓扑模式是研究蛋白的生物功能信息的重要的出发点之一，因此我们需要发展一些方法来挖掘和理解相互作用网络。这里我们把已经在其他领域成功应用的谱分析方法用到蛋白组的研究中，来识别蛋白相互作用网络中的拓扑模式，即准团(quasi-clique)和准二部图（quasi-bipartite）。有趣的是，我们发现在同一组的蛋白具有相似的蛋白功能。更重要的是对于酿酒酵母的近三分之一的未知功能的蛋白，这种方法提供了一种基于蛋白结构预测蛋白功能的手段。

谱分析是用来揭示海量复杂数据关系的深层结构的一种有效的方法。作为一个著名的范例，戴维.吉布森（David Gibson）、乔恩. 克莱因伯格（Jon Kleinberg）和普拉布哈卡尔.拉加瓦（Prabhakar Raghavan）在万维网（World Wide Web）链接结构的信息发掘领域做了出色的工作[80,81]。众所周知，万维网（World Wide Web）是由数量不断增加的网页通过超链接与其他网页链接而成。除了万维网结构的复杂性，谱分析方法还成功地发现了"权威点"

(authoritative)和“集成点”（hub）等数据信息源。我们把谱分析方法应用于复杂的蛋白一蛋白相互作用网络来识别有趣的拓扑结构。在这个方法中，网络被表示成一个无向的图 $G(V,E)$，也就是说节点集合包含每个蛋白作为节点：$V=\{P_1,P_2\cdots,P_n\}$，而边集合的定义为 $E=\{(P_i,P_j)|\text{具有相互作用的蛋白}P_i\text{和}P_j\}$。对称的 $n\times n$ 链接矩阵可以定义成 $A=(a_{ij})$，

$$a_{ij}=\begin{cases}1 & (P_i,P_j)\in E\\ 0 & (P_i,P_j)\notin E\end{cases}$$

链接矩阵 A 的谱本质上说是能通过相互作用传递的节点属性的一种重要的量度。可以给每个节点一个分数 x 来表示它的“重要性”。一个具有高分数的节点将通过相互作用增加和它相链节点的得分，即两个相互作用节点的得分相互加强，可通过循环定义它们的得分

$$\varDelta X_i=\sum_{j=1}^{n}a_{ij}\times X_j$$

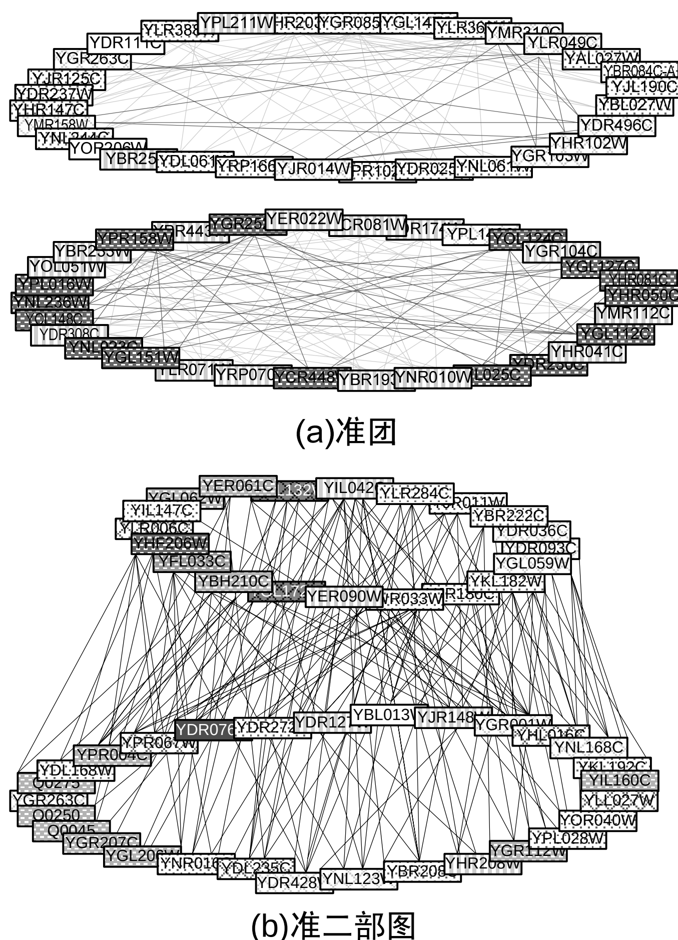


图5. 蛋白质相互作用网络中的准团与准二部图

吉布森等人的迭代算法介绍了一种中断这种循环的方法。有趣的是无论任何的初值情况下，X_i 都将聚集到一个特定的点。可以证明这个点正是矩阵 A 的一个本征向量。这就证明其性质是相互作用的一个本质属性。不仅如此，因为 A 矩阵是对称矩阵，所有的本征向量都是正交的。这就意味着可能对应的属性也是正交的。换种说法就是每个本征向量可能表示了一种特别的其它向量没有的性质。从拓扑的观点看，图的谱帮助揭示复杂相互作用网络的隐含结构。我们发现对于每个对应一个正的本征值的本征向量，其绝对值较大的分量倾向于形成一个准团（Quasi-cliques，即正负两端分别形成一个趋于内部链接的集团）（见图 5a），而对于每个负本征值的本征向量，这样的蛋白趋向于形成准二部图（Quasi-bipartite，即正负两端的内部不相连的蛋白形成一个趋于紧密相链的结构）（图 5b）。

在高通量方法产生的相互作用数据中有大量的假阳性存在。为了度量这些数据的精确度和确定其偏歧，冯. 梅林（von Mering）等评估了已经发表的 5,400 多蛋白中的 80,000 多条相互作用，给每个相互作用一个置信度打分[82]。为降低数据的假阳性，我们把对数据的分析集中在中、高置信度的数据上，包括了 2,617 个蛋白的 11,855 个相互作用。为了分析相互作用数据，首先，我们把谱分析方法应用于计算对应网络的邻接矩阵的所有本征

值和本征向量。用以下的标准在大的正的本征值的本征向量上产生准团：（1）所有的蛋白都按照其本征值的绝对值进行排序，先选取 10%的本征向量进行分析；（2）按照排序的顺序加入蛋白，新加入的蛋白至少与已有的蛋白有 20%部分有相互作用。这里我们用成团系数 CC 来衡量节点之间的连接关系的程度，调节参数以保证准团的性质；（3）准团至少要包含 10 个蛋白。按照这样的标准取团，我们得到了 48 个准团。其中最大的一个包含 109 个蛋白，而最小的一个包含 10 个蛋白，平均具有 26.6 个蛋白（一个蛋白可以出现在多个团中）。相似的分析可以用在对负值的本征向量的分析中，得到 6 个准二部图。这两个拓扑图谱显示了不同的相互作用谱。在准团中，蛋白倾向于与自己相互作用（见图 5a），而在准二部图（quasi-bipartite）中两个集合间趋向于有相互作用，而其内部没有相互作用（见图 5b）。对这两种模式的识别不仅使得对复杂相互作用网络的表示更有序化，更重要的是，提供了一种能够更方便分析复杂网络的手段。一个孤立的准团包括不同的生物功能。P 值的方法可以作为一个给准团赋予主要功能的标准。超几何分布可以计算对应蛋白团对于某个功能的概率。对于蛋白数为 n 的团，含有某功能蛋白 k 个，设其所在的蛋白组共有 G 个蛋白，该功能类共有 C 个蛋白，这样的团随机出现几率的 P 值是：

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-c}{n-i}}{\binom{G}{n}}$$

上面的标准描述了蛋白准团具有某个特定功能类的丰度与随机出现相比的几率。如果 P 值的丰度接近于零，表示准团中这类蛋白随机选出的几率会很低。这里我们把每个准团所有功能类中 P 值最低的功能作为该团的主要功能。对于 48 个准团中的每个团，我们都采用慕尼黑信息中心（Munich Information Center，MIPS）的层次功能注释对其进行了注释，并计算了功能注释的 P 值。在计算 P 值过程中，MIPS 的注释允许一个蛋白有多于一种的功能。结果显示其中的 43 个准团都可以被赋予一种功能，而其他的 5 个准团可以被赋予几种功能。对准团中的单个蛋白的功能分析研究发现大部分的蛋白趋向于同一共有的功能，比如核糖体的生物起源（ribosome biogenesis），核糖体核糖核酸和转移核糖核酸的合成（rRNA and tRNA synthesis）、处理(processing,) 、转录调控(transcription control)和信使核糖核酸剪切(mRNA splicing)等。只有一小部分的蛋白是没有标注功能的或者是具有和准团中主要功能相冲突的功能，如（图 6）所示。

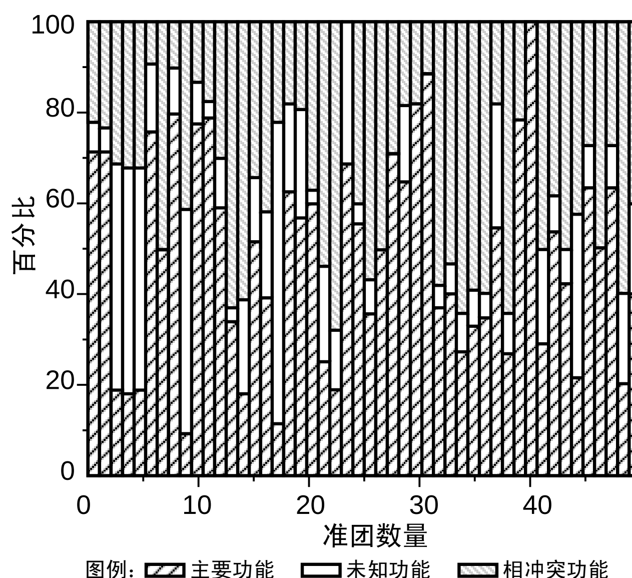


图6. 48 个准团的功能分类的百分比

分离出来的准团为预测没有标注功能的蛋白的功能提供了很好的线索。在 2,617 个蛋白的原始数据中，有 555 个蛋白在 MIPS 的层次功能分类中是没有标注的。对于 48 个准团中，包含有 76 个没有标注的蛋白。我们对每个这样的蛋白用它们所在团的主要功能对它们的功

能进行预测。如果一个蛋白落在了多个团中，就用具有最小 P 值的团进行预测；如果多个团都具有最小 P 值而团又具有多个功能，则赋予未知蛋白多个功能。其中有 43 个蛋白是和核糖体核糖核酸处理(rRNA processing)有关；7 个和核糖核酸前端处理（pre-RNA processing）有关；11 个蛋白与核糖体生物起源（ribosome biogenesis）有关；其它 15 个蛋白分别与能量(energy)、代谢（metabolism）、细胞骨架（cytoskeleton）和转录调控（transcription-regulating）有关。我们用吴（音译，Lani F. Wu）等的计算 P 值来对功能标注的方法进行评估。作为对照，我们产生并分析了与原网络具有相同的度分布的随机的网络数据。结果显示，对于我们的实验数据分析中的 48 个团中有多于 87.5%的功能类的注释是有意义的（即 $p \le 0.01/N_c$，这里 N_c 是功能类的总数），而随机网络产生的准团的对照组中，只有 2.1%的准团满足这样的标准。这就意味着分离出的准团有相当一部分可能具有生物意义。我们对准团的一些预测结果已经得到了最近的实验证据的证明。在所有的这些准团中，有五个是被未标识的蛋白所占据的（即未知蛋白至少占所有蛋白数目的 50%）。这就暗示着这些在同一个准团下的未知蛋白可能组成一个与某个特定细胞过程相关的复合体。如图 7 所示，在我们预测的准团中，根据我们的预测大部分蛋白都与 rRNA 处理相关，这与最近的实验结果正好部分地吻合。

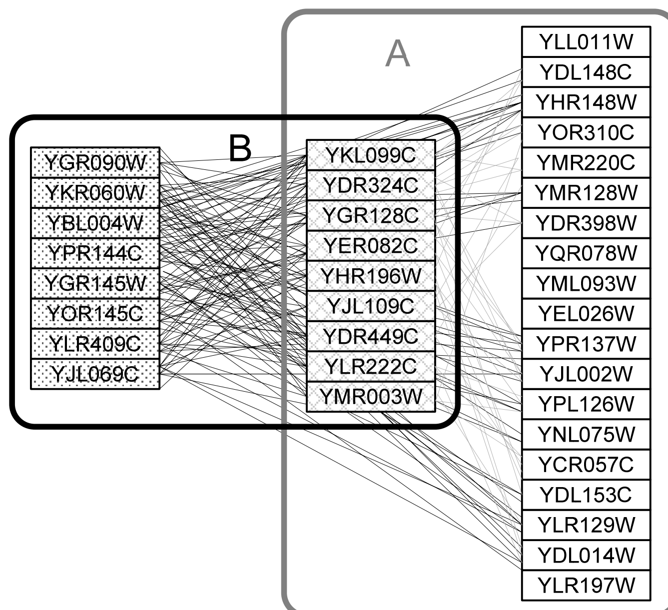


图7. 功能预测和小亚基处理复合物注释实验结果的比较

3.2 转录调控网络的调控模式倾向性分析

转录调控网络控制着细胞中所有基因的表达水平，对转录调控网络的研究是后基因组时代的一个重要问题。随着相关实验技术的快速进展，现在已经有多个模式生物的转录调控网络经实验测出[83,84]。我们可以把转录调控网络简单地看作是一个有向图。在这个图里，转录因子（transcription factors，TFs）和转录因子所调控的基因（transcription target genes，TGs）表示为图中的节点。而转录因子对它所调控的基因的调控作用就是转录因子绑定到它目标基因的上游转录调控区从而控制该基因的转录。在图中就表示为从转录因子到该被调控基因的一条有向边。转录因子和它们所调控的基因之间的调控关系在图中就表现为一个具有多个点的子图。有些子图因其拓扑上具有明显的生物学含义而被广泛地研究[83-85]，比如前馈环（feed-forward loops）、反馈环（feedback loops）、单输入模体（single input motifs）和多输入模体（multi-input motif）（见图 8）。这些子图，或者说调控模式可能含有特定的调控能力。比如单输入模体可能用来调节一组功能相关的基因；而前馈环则有可能在某个生物过程中起到时间控制的作用。但是这些子图在网络中并不是和网络的其他部分没有关联的一个个独立的功能单元。事实上，这些子图倾向于在那些具有很高连通性的网络中心转录因子周围聚集。这样，一个转录因子往往成为多个不同模式子图中的成员。

从全局上来说，调控网络的拓扑结构分析表明，转录因子所调控的基因数目服从一种幂律（Power-Law）分布。这意味着，在转录调控网络中一小部分的转录因子调控了大多数的

基因。这些具有高度连通性的转录因子被称为网络中心转录因子（transcription hubs，THubs）。研究表明，这些网络中心转录因子通常在生物体中是至关重要的关键基因[86-88]。转录调控网络可以看作一个用于信号（比如外界营养物质，环境压力）传递的网络结构，而不同转录因子在网络中传递信号时表现出的行为应该存在差异[89]。类似情况最近在哺乳动物的信号转导网络中发现，在网络不同层次三个重要的配体采用了不同的子图发挥功能[90]。另一方面，在不同的外界条件或者发育时期的转录调控子网中，不同的调控模式在网络中总的密度也存在变化[91]。但是，到目前为止，还没有人在基因组的尺度上来了解这种不同的模式密度对于每一个转录因子的影响；也没有方法对转录因子对其下游调控基因进行转录调控时的行为进行度量。为此，我们希望能够设计一套方法来测度和表示转录因子在调控网络中对不同调控模式的使用偏好。

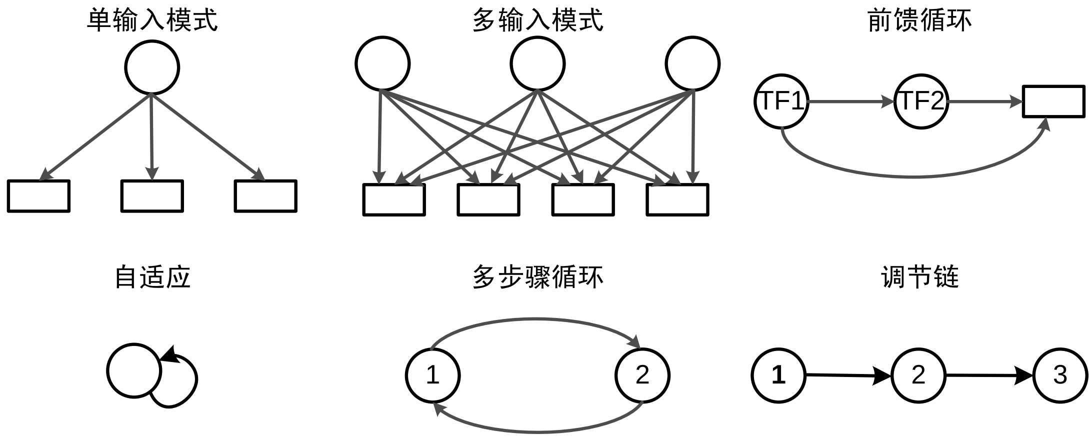


图8. 转录调控网络中的模体示意图

为了计算网络子图倾向性谱，首先需要决定我们研究的转录因子集合和子图集合。我们研究的转录因子主要集中在中心转录因子集合上。对于度分布服从幂律分布的复杂网络，如

转录方法模式	方法
3顶点环型模式	（R3-1）（R3-2）
3顶点树型模式	（T3-1）（T3-2）（T3-3）
4顶点环型模式	（R4-1）（R4-2）（R4-3）（R4-4）
4顶点树型模式	（T4-1）（T4-2）（T4-3）（T4-4） （T4-5）（T4-6）（T4-7）（T4-8）

图9. 环形模式和树型模式两类基本调控模式

转录调控网络，其中多数节点只与很少的其他节点连接，而在总数中很少的节点却连接了大

量的其他节点。这样的节点，被称为中心节点（Hub）。在转录调控网络中我们称这些中心节点为中心转录因子，它们调控的基因明显高于转录网络中的其他大部分转录因子。这些中心转录因子通常在生物体中是关键基因。中心节点的确定通常是以在节点的分布曲线中取拐点作为阈值，连接度大于这个阈值的节点被定义为中心节点。对于基本调控模式，我们选择了环形模式和树型模式两类（见图 9）。

我们判定一个调控模式的类型是基于其在拓扑上的特性是开放的还是封闭的。一个调控模式不管其调控方向如何，如果在拓扑上构成一个单独的封闭环，则这种调控方式被称为“环”。一个调控模式不包含任何环作为自己的子图，则被称为“树”。我们只考虑三个节点和四个节点的环和树，这是因为在两个节点的层次，T2-1 就是平凡的调控结构，而 R2-1 则是在网络中非常罕见的模式。而对于节点数高于 4 个的调控模式因为计算能力上的限制，也没有包括在我们的研究范围内。我们之所以选择这些作为基本调控模式不仅是因为其相互间的相似性足够小，而且也是因为这些调控模式涵盖了所有的基本模式，即所有其他的模式都可以从这些基本模式的组合中生成出来。 选定转录因子集合和子图集合以后， 对每一对给定的转录因子 H 和子图 P 我们定义这个转录因子对于这个子图的使用倾向性 A_w 如下：

$$Aw(H,P)=\sum_{sg\in SG(H,P)}\sum_{k\in N(sg)}1/(d(H,k)+1)^2$$

其中 $SG(H,P)$ 是转录因子 H 下游所有子图 P 的实例的集合， $N(sg)$ 是子图实例 sg 中的所有节点， $d(H,k)$ 是转录因子 H 和其下游基因 k 在网络中的最短距离。公式中出现的权重因子， $1/(d(H,k)+1)^2$ ，用来量化一个转录因子对下游的影响随着距离的增加而逐渐减小的特点。

对于选定的转录因子集合和子图集合，每一对转录因子和子图都可以计算出转录因子对于子图的使用倾向性。不过由于不同的转录因子所调控的下游区域的规模是不一样的，同时，不同的子图在网络中的总体丰度也有很大差异，所以公式得到的这些不同转录因子和子图之间的使用倾向性值不能直接比较。为了让这些使用倾向性值能够互相比较我们采用如下公式来消除上述的两个因素的影响：

$$A_N(H_i,P_j)=\frac{A_w(H_i,P_j)/\sum_{j\in SG}A_w(H_i,P_j)}{\sum_{i\in THubs}A_w(H_i,P_j)/\sum_{i\in THubs}\sum_{j\in SG}A_w(H_i,P_j)}$$

其中 $A_N(H_i,P_j)$ 是归一化之后的转录因子 H 对子图 P 的倾向性，$THubs$ 代表我们研究的转录因子的集合，这里是全部的中心转录因子，SG 代表我们研究的全部基本调控模式（三个节点或四个节点的所有环形模式和树型模式）。这样我们就得到了所有中心转录因子对所有子图的“归一化的使用倾向性”。我们称一个转录调控因子的“网络子图倾向性谱（subgraph preference profile，SPP）”为由这个转录因子对所有子图归一化的使用倾向性值所构成的一个向量。而我们称所有待研究的转录因子的网络子图倾向性谱构成的矩阵为这个网络的网络子图倾向性蓝图（subgraph preference landscape，SPL）。

我们可以把子图倾向性蓝图可视化为一个灰度图（如图 10）。在图中显示得非常黑的点表示转录因子对这些调控模式不同于其他调控模式的倾向性。但是因为有研究显示，网络的全局性结构和某些局部模式是相互决定的，所以，我们还不能从这个灰度简单地判定转录因子是真的倾向于使用这个调控模式还是只不过是具有这类丰度分布的网络所通有的特性。为了给出一个网络子图倾向性的显著性判定，对于一个特定丰度分布的具体网络，我们考察从

这个网络出发随机生成的随机网络簇，使得生成的随机网络和真实网络具有相同的出度和入度的分布。通过计算在随机网络簇中的网络子图倾向性蓝图，我们可以得到调控模式倾向性在随机网络簇中的分布。因为转录调控网络的随机网络子图倾向性值大致服从幂律分布，我们采用一种基于 z 分数法（z-score，到均值距离相对标准差的倍数)的算法来确定倾向使用某一子图的显著性阈值：从随机网络簇中剔除那些子图倾向性值的 z 得分大于等于 2 的数据，然后根据剩下来的随机网络簇的子图倾向性值分布，我们取 z 得分大于等于 2 作为用以判定显著倾向性所使用的阈值，那些子图倾向性大于这个阈值的调控模式是被所对应转录因子显著倾向性使用的。为了给出这个倾向性的统计显著性， 我们把真实网络和 1000 个随机生成的网络进行比较，每一对转录因子和网络子图的倾向性的 p 值由这个倾向性值小于在 1000 随机网络中对应的倾向性值的百分比给出。

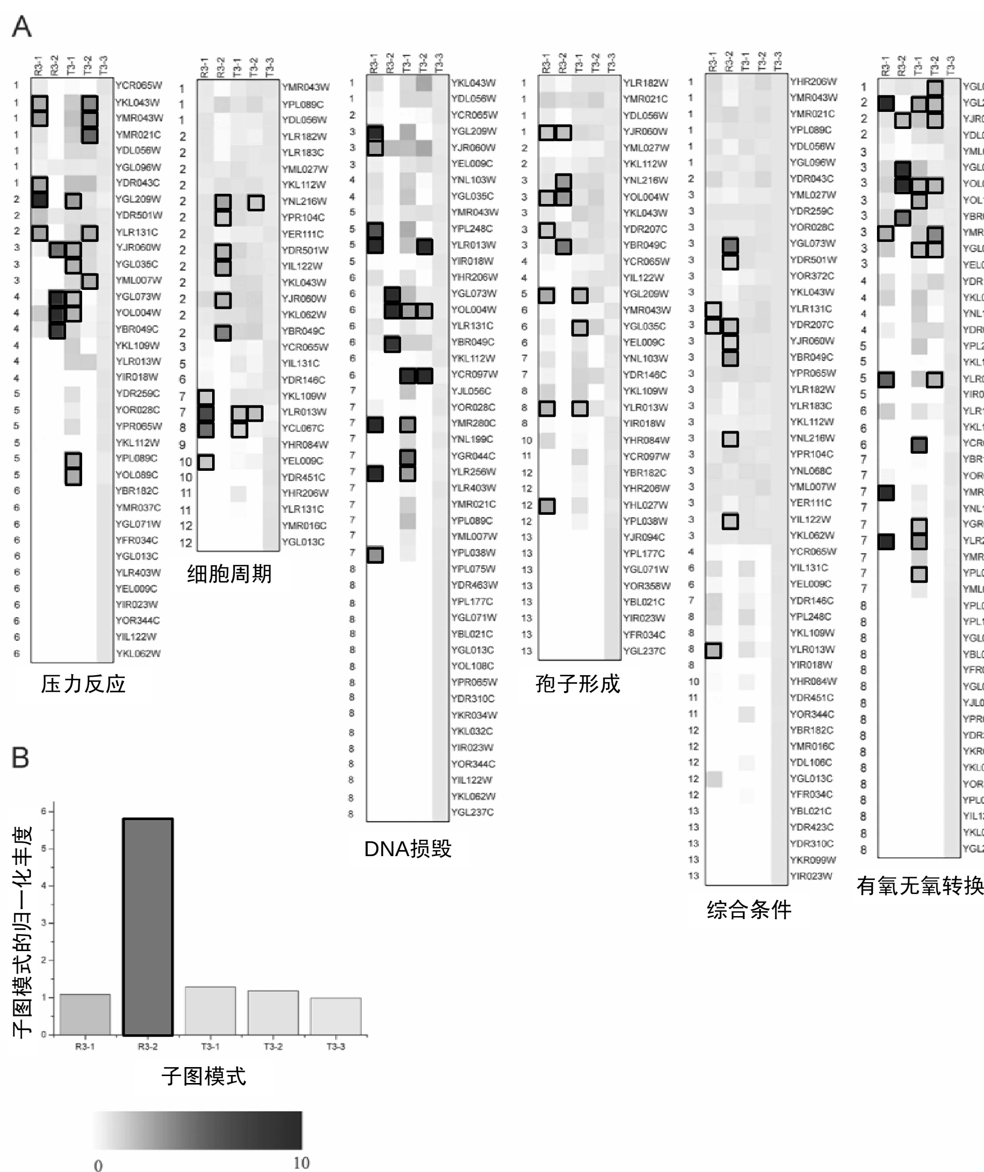


图10. 啤酒酵母转录调控网络的网络子图倾向性谱和网络子图倾向性蓝图

对于一个基于网络的定量计算分析方法，我们除了要考虑结果相对于随机网络是否有显著性意义，还需要考虑结果是否具有一定程度的鲁棒性、稳定性。特别是考虑到转录调控网络的特殊性， 高通量实验带来的各种噪音比较大， 这一个问题就显得尤为重要。对于鲁棒

性分析，我们从下面三个方面来考察：

1. **对网络节点规模的鲁棒性** 我们在网络中随机的敲除和添加一些节点。然后观察添加/删除节点对网络子图倾向性蓝图的影响。
2. **对噪音的鲁棒性** 我们给网络中加入一定的连接噪音，这种噪音包括随机地加入、删除和交换一些连接。 然后观察噪音对网络子图倾向性蓝图的影响。
3. **对下游基因数目的鲁棒性** 我们改变原来的阈值，重新得到新的转录中心因子集合，然后观察对网络子图倾向性蓝图的影响。

此外，在我们研究的网络中，对中心转录因子的认定是采用它们度分布的拐点作为阈值。这个阈值的选取没有绝对的标准，为此我们需要考察阈值的选定对子图倾向性蓝图的影响力。我们对每一个网络把选定的阈值分别增加和减少 1，分别得到两组新的中心转录因子集合。利用这些不同的中心转录因子集合，我们重新计算了它们的子图倾向性蓝图。对于以上三种情况，我们分别产生了相对于原始网络的参照网络簇，然后计算了这些网络的子图倾向性蓝图。对每一个给定的原始网络和与之对应的网络簇我们通过比较它们的子图倾向性蓝图来判定其对于相应操作的鲁棒性。比较方法为：对每一对子图倾向性蓝图（原始网络和对应网络簇中的一个网络）计算其对应的所有网络子图倾向性谱中向量对之间的欧氏距离，得到一个网络子图倾向性谱间距离的分布；根据参照网络簇内部子图倾向性蓝图之间对应网络子图倾向性谱距离分布可以得到网络子图倾向性谱之间是否存在显著差异的阈值；根据阈值我们可以判定原始网络子图倾向性蓝图和对应网络簇子图倾向性蓝图间的距离存在显著性差异的 p 值。通过上述鲁棒性分析我们发现网络子图倾向性谱和网络子图倾向性蓝图方法的分析结果对于各种点、边的噪音以及转录中心因子定义阈值的改变具有很好的稳定性。

进一步，我们考察了啤酒酵母的转录调控网络中心转录因子和它们下游调控模式的倾向性关系。我们的分析包括了综合条件（static）、细胞周期（cell cycle）、孢子形成（sporulation）、有氧无氧转换（diauxic shift）、DNA 损坏（DNA damage）和压力反应（stress response）共 6 个条件下的啤酒酵母转录调控网络。其中综合条件网络是全网，另外五个网络是全网在各种条件下的子网。网络数据来自 http://sandy.topnet.gersteinlab.org/，网络中的自相互作用边被去除。和先前在大肠杆菌中所发现的类似，酵母的转录调控网络也是一个多层的层次结构。在酵母的综合条件调控网络中一共有 14 层，而在细胞周期，孢子形成，有氧无氧转换，DNA 损坏和压力反应这五个条件下的子网络中分别有 13，14，9，9 和 7 层。当我们把子图倾向性蓝图中中心转录因子的网络子图倾向性谱按照他们在层次结构中的顺序排列的时候，我们在所有的子图倾向性蓝图中都观察到一个普遍的倾向：在网络偏上部分的转录因子比网络偏下部分的转录因子有着更复杂的网络子图倾向性谱（图 10）。

进一步，我们先分析了在各种条件下的子图倾向性蓝图的特点，然后对不同条件下的子图倾向性蓝图进行了比较分析。图 10 给出了各个网络的调控模式倾向性蓝图，其中每个方格中的灰度反映的是偏好程度，越深偏好程度越高，显著性高的部分我们用方框表示出来。我们可以明显地看到在调控网络中不同转录因子有倾向性地使用不同的调控模式。 我们还注意到单输入模体（T3-3 和 T4-7）虽然是唯一在所有的网络中和在所有的层次中都出现的一种调控模式，但是在所有的 6 个转录调控网络中没有任何转录因子倾向于使用这种调控模式。而对于前馈环（R3-1）则在网络的各种层次上都存在转录因子倾向于使用这种调控模式。在细胞周期和孢子形成这两个子网中，反馈环（R3-2，R4-1）被网络高层的转录因子倾向性地使用，而在另外的其他三个子网（有氧无氧转换，DNA 损坏和压力反应）中，更多是在网络下层的转录因子倾向于使用这种调控模式。有些调控模式在网络中的相对高丰度不能用来解释转录因子对这些调控模式的倾向性使用。例如在调控网络中研究发现前馈环、反馈环、

单输入模体是相对高频度出现的调控模式，被称之为网络的模体（network motif）。但是，如我们上面提到的，单输入模体在我们考察的所有网络中没有被任何转录因子倾向性地使用，相反，有些并不是很显著高频出现的调控模式，比如 T3-1，T3-2 却被某些转录因子倾向性地使用了。转录因子对某类调控模式的高显著性使用有可能是调控模式在该因子周围聚集的结果。例如，在细胞周期的调控网络子网中的转录因子 YLR013W 高度倾向性的使用前馈环。仔细考察 YLR013W 在调控网络中的上下游，我们看到有 4 个前馈环形成一个对称的网格形式（见图 11 (a)）。但是并不是所有的高倾向性使用都可以用聚集来解释。在综合型的网络中，我们总共探测到 3 例反馈环（R3-2），这些反馈环不仅聚集在一起，而且前后连接在一起成为一个大的反馈环，并且在这个大的反馈环中的所有节点都是中心转录因子（见图 11(b)）。然而，尽管如此，在这个大的反馈环中仍然存在不倾向性使用反馈环的中心转录因子（如 YGL073W），并且在所有网络中都倾向性使用了反馈环的中心转录因子只有一个 YBR049C。既然转录因子对调控模式的倾向性使用既不能由这些模式的高丰度完全解释，也不能被这些模式在局部区域的高度聚集来完全解释，我们认为这暗示了我们定义的这种"倾向性使用"是反映了这些中心转录因子在一定生长或者细胞条件下的转录调控网络中的某种重要行为偏好性。

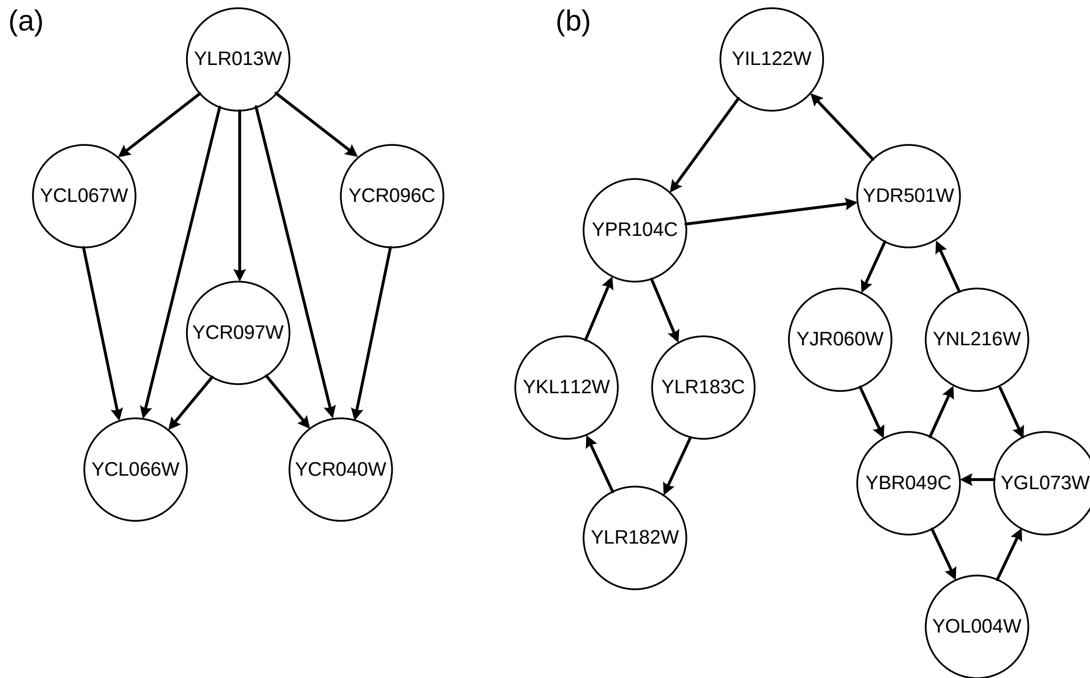


图11. 对前馈环、反馈环倾向性使用的示例

当我们将五个不同条件下的子网的网络子图倾向性蓝图进行比较时，观察到了不同网络在对调控模式的倾向性上的动态变化。首先我们用科尔莫戈罗夫-斯米尔诺夫（Kolmogorov-Smirnov）检验来测试在不同条件下子网的调控模式倾向性蓝图中的倾向值是不是来自同一分布。在有氧无氧转换、DNA 损毁和压力反应的子网（它们又被称为外源性网络）的三个节点，子图的倾向性蓝图明显地区别于内源性网络(细胞周期和孢子形成)（见表 2）。但是在孢子形成的子网和有氧无氧转换条件下的子网之间有些类似，这个例外也许是反映了孢子形成过程中的某些外源性特性的影响。在外源性网络之间，我们看到其子图倾向性蓝图基本上是来自同一个分布，相反在两个内源网络之间，他们的调控模式倾向性蓝图中的倾向性 p 值分布是明显不同的。在四个节点的子图倾向性谱中我们也观察到了类似情况。其次，通过比较同一层内转录因子的子图倾向性谱之间的欧氏距离，我们观察到在不同的调控网络中，

表2. 对 3 节点网络子图倾向性值分布的科尔莫戈罗夫-斯米尔诺夫检验

条件	综合	细胞周期	孢子形成	DNA 损毁	有氧无氧转换	压力反应
综合	——	(0.1931)[a]	(0.0086)[b]	$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]
细胞周期	(0.1931)[a]	——		$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]
孢子形成	(0.0086)[b]	(0.0071)[b]	——	(0.0133)[b]	(0.0063)[b]	(0.1611)[a]
DNA 损毁	$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]	(0.0133)[b]	——		(0.0905)[a]
有氧无氧转换	$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]	(0.0063)[b]	(0.3438)[a]	——	(0.1294)[a]
压力反应	$(<1\times10^{-10})$[b]	$(<1\times10^{-10})$[b]	(0.1611)[a]	(0.0905)[a]	(0.1294)[a]	——

同一层内子图倾向性谱间的相似性是不一样的。对于三个节点的调控模式来说，在综合性网络、细胞周期和压力反应三个网络中处于同一层的转录因子通常倾向于有更相似的网络子图倾向性谱（见表 3），而在其他三个网络中，处在同一层中的转录因子则倾向于有不同的网络子图倾向性谱。在四个节点的层次，在综合条件网络、细胞周期子网和有氧无氧转换子网三个网络中处于同一层的转录因子通常倾向于有更相似的模式倾向性谱，而在其他三个网络中处在同一层中的转录因子则倾向于有更加不同的倾向性谱（见表 4）。因为对于底层的转录因子来说他们的倾向性谱是简单而且明显的，即只包含单输入模体，所以为了去除这种明

表3. 同一层内部 3 节点网络子图倾向性谱间相似性比较

条件	全部	Round All	*p*-值	内层	Round Inner	*p*-值
综合	1.354	1.697	<0.001	1.405	1.649	<0.001
细胞周期	1.972	2.511	0.006	1.995	2.489	0.012
孢子形成	1.17	2.233	<0.001	2.227	2.394	0.226
DNA 损毁	3.461	6.3	<0.002	8.235	8.388	0.434
有氧无氧转换	3.299	5.234	<0.003	8.127	6.948	0.913
压力反应	2.853	5.743	<0.004	5.605	7.214	0.042

表4. 同一层内部 4 节点网络子图倾向性谱间相似性比较

条件	全部	Round All	*p*-值	内层	Round Inner	*p*-值
综合	1.474	2.799	0.002	1.529	2.725	<0.001
细胞周期	3.172	6.129	<0.000	3.21	10.688	0.161
孢子形成	2.943	2.799	<0.001	5.601	7.058	<0.001
DNA 损毁	20.061	37.368	<0.002	47.731	54.405	0.322
有氧无氧转换	4.984	10.638	<0.003	12.278	14.409	0.092
压力反应	12.493	21.148	<0.004	24.548	27.669	0.195

显的偏差影响，我们在做层内的模式倾向性相似分析的时候不包括最底层的转录因子。最后我们考察在五个条件下的子网中都被认定是中心转录调控因子的九个转录因子的动态特性（见图 12）。对于三个节点的调控模式来说，在这九个转录因子中，只有 YLR013W 的倾向性谱在两个内源性网络之间有显著的变化，尽管两个内源性网络的倾向性的分布是完全不同的。相反的是，在三个外源性网络之间却有四个中心转录因子（YMR043W，YJR060W，YKL043W，YLR013W）的倾向性谱有着显著的变化，尽管这三个外源性网络的倾向性的分

布是相似的。在四个节点的层次我们可以观察到倾向性谱的更多的动态变化。

这些动态变化也许反映的是转录因子在外界环境或者生长状态的变化过程中所行使的生物功能的动态转化。例如，在细胞周期和压力反应的转录调控网络中，核小体行使功能所需要的 YJR060W 倾向性地使用了三节点和四节点的反馈环[92]，而在 DNA 损毁的转录调控网络中 YJR060W 则转为倾向性地使用前馈环[93]。作为同时钟和震荡器类似的基因调控模式，反馈环也许是调控着细胞的生长速率。因此，在细胞周期和压力下这个转录因子倾向性地使用反馈环暗示了也许它的作用是一个时钟控制器或者频率调节器，而在 DNA 损毁的过程中，YJR060W 则有可能起到一个信号放大器的作用。这个例子说明，比较子图倾向性蓝图在不同条件下的变化可以给出转录因子对不同外界环境行使不同生物学功能的隐含意义。

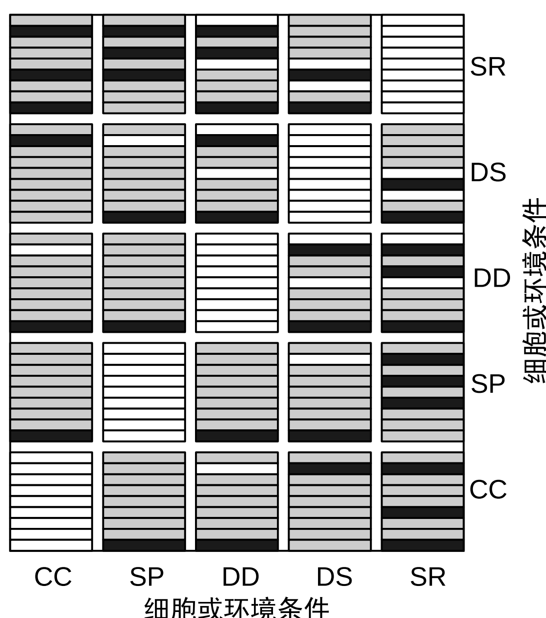


图12. 九个转录因子的动态特性分析（3 节点子图倾向性谱）

4　非编码 RNA 参与的复杂生物网络

生命体绝大多数活动都涉及很多生物分子（包括基因、蛋白质、及其他生物分子）的复杂相互作用。这些相互作用关系实际上表现为网络的关系，而生物学功能在这种网络相互作用中体现出来。换言之，生物体的复杂性不仅源自其内部大分子及其种类的巨大数目，更因为大分子之间错综复杂的关系。复杂网络表现为三个层次上的相互作用关系。其一是基因转录调控网络，表现为基因转录过程中转录因子和其调控基因上游的转录因子结合位点的相互作用。其次是蛋白质相互作用网络。再次是转录后调控网络，特别是微型核糖核酸和其调控序列的相互作用。大量非编码核糖核酸的加入对于生物网络的复杂性影响巨大。首先，已有研究表明在高等生物中非编码核糖核酸数量庞大，ENCODE 项目的研究结果表明人类基因组的 93%都会转录成核糖核酸，其中超过一半都是非编码核糖核酸[94]。因此非编码核糖核酸的加入将使网络规模成倍的扩张。另外，更重要的是非编码核糖核酸的加入增加了各种新的相互作用机制，对于基因的转录调控，以及转录后调控和修饰都发挥着重要作用。例如微型核糖核酸，这种广泛存在于高等动物和植物中的微小的非编码核糖核酸，通过控制信使核糖核酸的稳定性或抑制信使核糖核酸的翻译对生命活动起到重要调控作用[10]。由于认识到非编码核糖核酸的出现对生物网络带来的重大影响，对非编码核糖核酸参与的复杂生物网络的研究也已经启动。将非编码核糖核酸加入生物网络研究有助于我们更好地了解生物网络，大大地丰富了我们对整个生物网络的认识。非编码核糖核酸可以与特定蛋白质相互作用形成各种复合物，以核糖核酸-蛋白质复合物的形式行使其功能。如 snRNA U1、U2、U4、U5、U6 同多达 75 种蛋白质组成剪接复合物，负责信使核糖核酸前体（pre-mRNA）的剪接[29]；小鼠的 NRON RNA 同 11 种蛋白质结合，控制 NFAT 蛋白的转运 [95]。非编码核糖核酸还可以通过与靶核糖核酸（target RNA）序列匹配来定位靶标，并进一步招募功能蛋白质来行使功能。如 C/D box snoRNA 通过其上的互补序列定位核糖体核糖核酸上的作用位点，并招募

蛋白质来对核糖体核糖核酸进行甲基化修饰[25]；微型核糖核酸通过其种子（seed）序列定位于特定信使核糖核酸的 3' UTR[5]区域，并通过其招募的诱导沉默复合体（RNA induced silencing complex，RISC）蛋白来控制信使核糖核酸的稳定性或抑制信使核糖核酸的翻译[10]。将非编码核糖核酸加入网络来进行研究还有助于我们研究非编码核糖核酸本身的功能。在对于编码基因的研究中，蛋白质相互作用网络以及基因转录调控网络的研究已经展现了网络研究的巨大威力：通过网络聚类寻找功能模块，根据网络邻居节点预测蛋白质功能。这些基于网络的研究已经成为生物学研究的新武器。我们相信这些在网络研究中已经证明非常成功的分析方法也肯定能够帮助我们更好地预测非编码核糖核酸的功能。

微型核糖核酸是一种广泛存在于高等动物和植物中的微小的非编码基因，通过控制信使核糖核酸的稳定性或抑制信使核糖核酸的翻译对生命活动起到重要调控作用。微型核糖核酸从染色体上转录出来的初级转录本（pri-miRNA）在细胞核中经核糖核酸酶 Drosha 处理后变成了微型核糖核酸前体（pre-miRNA）。然后微型核糖核酸前体被 Exportin-5/Ran-GTP 运送到细胞质中，在胞质中前体进一步被核糖核酸酶 Dicer 剪切成约 22 碱基对（bp）的双体（miRNA duplex）。这个双体将被一种核糖核酸酶解开，其中一条链将和蛋白质结合形成核糖核酸诱导沉默复合体（RISC），通过和信使核糖核酸 3' UTR 部分地互补配对来抑制蛋白质的合成或是对靶基因降解。一些研究组基于已知的微型核糖核酸调控特征对其靶标进行了大规模的预测，结果显示人类基因组中有近 1/3 的基因受到微型核糖核酸转录后水平的调控，每个微型核糖核酸平均调控了数百个编码基因[10]。众多研究者正在力图通过微型核糖核酸和相应的靶基因建立转录后的调控网络，并通过研究微型核糖核酸转录后调控网络来研究微型核糖核酸与相应靶基因的生物功能。在已有的研究中所发现的微型核糖核酸靶基因都是编码蛋白质的基因，即信使核糖核酸。我们猜测微型核糖核酸可能可以调控一类特殊的非编码核糖核酸信使核糖核酸样非编码核糖核酸的转录水平，形成一个对非编码核糖核酸的转录后调控网络（见图 13）。大部分已知的非编码基因的长度都是比较短的，但是最近几年的几个重要模式生物的全基因组芯片实验和全长互补脱氧核糖核酸文库建设都发现，基因组上存在

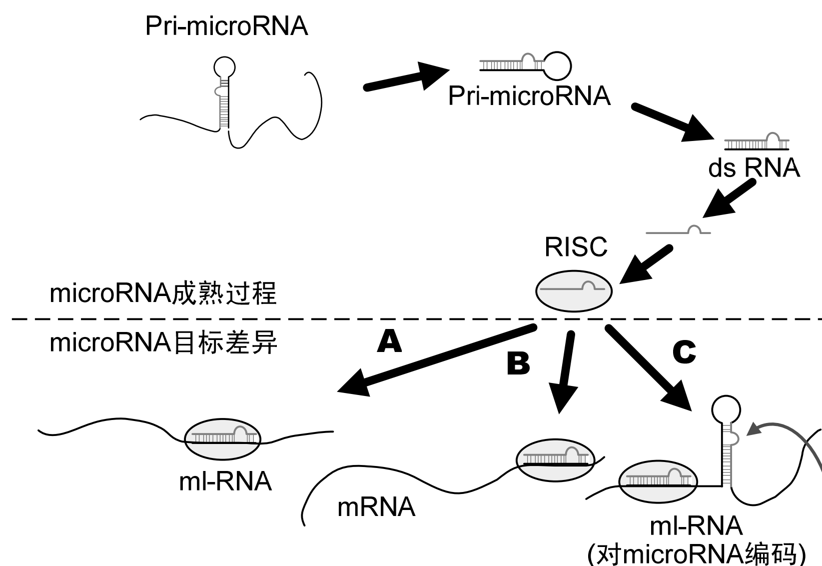


图13. 微型核糖核酸转录后调控关系示意图

着大量的长的非编码转录本。它们和编码蛋白的信使核糖核酸有一些相似之处：长度都很长，都由核糖核酸聚合酶 II 转录，转录后都存在剪接、加帽加尾现象，但是又没有蛋白编码框，

[5] Untranslated Regions，非翻译区，是信使核糖核酸分子两端的非编码片段

因此被称为信使核糖核酸样非编码基因[17]。少数信使核糖核酸样非编码基因的功能已经得到证实，然而绝大部分信使核糖核酸样非编码基因的功能和作用机制仍然是未知的。由于信使核糖核酸样非编码核糖核酸（mRNA-like ncRNA，mlRNA)同信使核糖核酸在序列和结构上的相似性，我们认为信使核糖核酸样非编码核糖核酸很可能也是微型核糖核酸的靶基因。为了验证我们的猜想，我们借鉴了已有对微型核糖核酸调控信使核糖核酸进行验证的方法[96]。近年来的研究表明微型核糖核酸能够加速其靶基因的核糖核酸降解，因此可以通过基因芯片检测所预测微型核糖核酸靶基因的核糖核酸水平来评估预测结果的可靠性。我们选取了 FANTOM 数据库中收集的 34000 条信使核糖核酸样非编码核糖核酸作为我们的研究对象。在这 34000 条序列中有约 11000 条序列在 20 个组织中有基因表达谱数据。我们又选取了 8 条已经确认的存在组织特异性表达的微型核糖核酸作为我们的微型核糖核酸研究集合（见表 5）。由于微型核糖核酸能够显著下调其靶基因的核糖核酸水平，所以对于组织特异

表5. 组织特异表达微型核糖核酸靶基因的表达谱分析

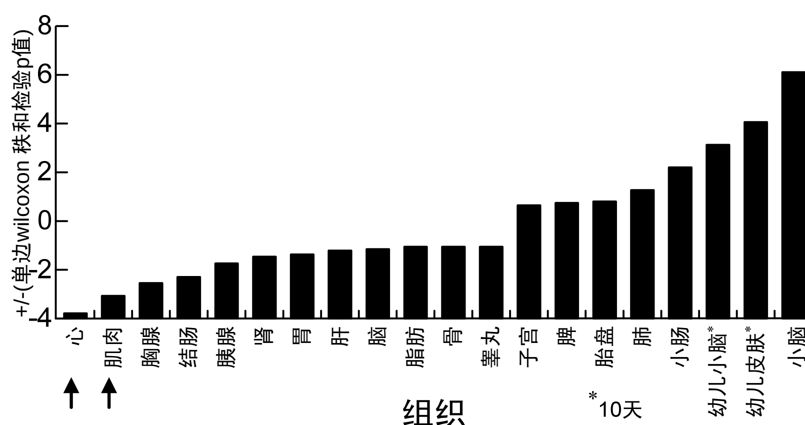
miRNA	组织	mlRNA 靶标		mRNA 靶标	
		排序	*P*	排序	*P*
miR-133a	心	1*	0.023	10	0.542
miR-133a	肌肉	2*	0.050	1*	0.005
miR-153	脑	5	0.254	18	0.944
miR-206	心	2*	0.007	3	0.116
miR-206	肌肉	3	0.071	2	0.056
miR-375	胰腺	2	0.168	12	0.648
miR-376a	胰腺	1*	0.0004	18	0.808
miR-122a	肝	16	0.714	2	0.0828
miR-124a	脑	12	0.458	1*	0.0007
miR-208	心	5	0.249	2	0.178

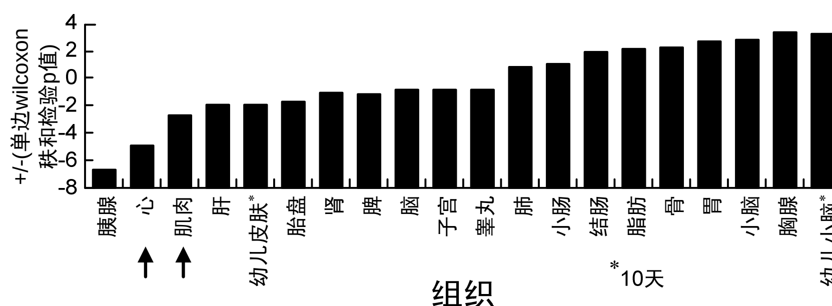
表达的微型核糖核酸，在其特异表达组织中其靶基因的表达水平应该显著低于其他组织。对于我们所分析的 8 条微型核糖核酸，我们根据 miRanda 预测结果以及靶位点序列保守性结果预测了它们在 11000 条信使核糖核酸样非编码核糖核酸集合上的靶基因。然后对预测靶基因的表达谱进行威尔科克森秩和检验（Wilcoxon’s rank sum test），结果如表 5 所示。结果表明有 3 条微型核糖核酸在 4 个特异表达的组织中其靶基因的表达水平显著下调（见图 14），显著水平和微型核糖核酸在信使核糖核酸上的调控水平相当，也验证了我们对于微型核糖核酸能够调控信使核糖核酸样非编码核糖核酸的猜想。我们的结果大大扩展了微型核糖核酸参与的转录后调控网络。更为有趣的是，我们的已有研究结果表明在信使核糖核酸样非编码核糖核酸中存在大量的微型核糖核酸编码核糖核酸（miRNA-encoding RNA），而微型核糖核酸对这些自身的初级转录本也存在调控关系，因此形成了一个复杂的微型核糖核酸间相互调控网络 （见图 15）。

5　结束语

本文结合近几年来我们研究组在非编码核糖核酸以及生物复杂网络方面所做的工作，分别对非编码核糖核酸功能研究、复杂生物网络分析以及非编码核糖核酸参与的生物网络的构建几个方面进行了介绍。将非编码核糖核酸引入网络研究是目前非编码核糖核酸研究以及生物复杂网络研究这两个生物信息学领域的热点问题交叉产生的一个前沿课题。在过去几年中这方面的研究虽然取得了一些成果，但目前仍然存在着很多有待解决的问题。已有的生物学

研究表明非编码核糖核酸对生物网络的影响是全局性的，参与到了基因转录调控、蛋白质相互作用以及基因转录后调控等网络的各个层次。而我们的工作现在还是局限于非编码核糖核酸参与的转录后调控网络这一层次。如何有效地研究非编码核糖核酸对生物复杂网络其他层次的影响仍然是一个问题。另一方面，对比传统的编码基因及其对应蛋白质构成的网络，新的网络将是一个混合的“双色网络”。在网络中存在双色节点（编码基因、非编码基因）。因此我们急需构建与分析这种非编码核糖核酸参与的双色网络的一个理论框架来指导我们的计算工作。将非编码核糖核酸引入生物复杂网络分析让我们得以从新的视角来研究生物学，也必然会带来各种新的问题和挑战。我们相信随着这些新问题和挑战的解决我们对生物学的理解也将进入一个新的阶段。





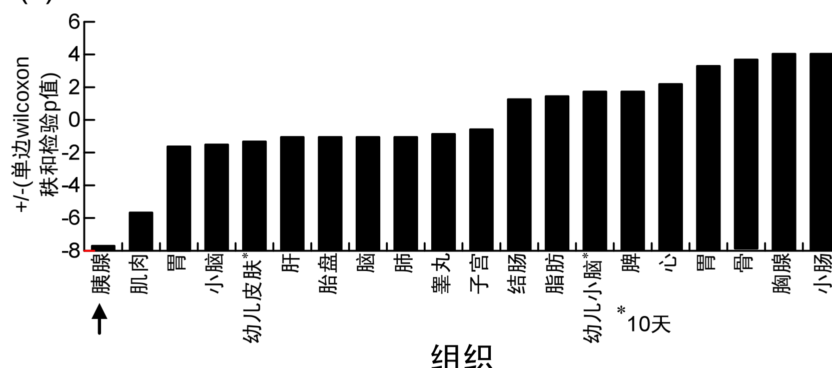


图14. 组织特异表达微型核糖核酸及其靶基因显著下调的对应组织

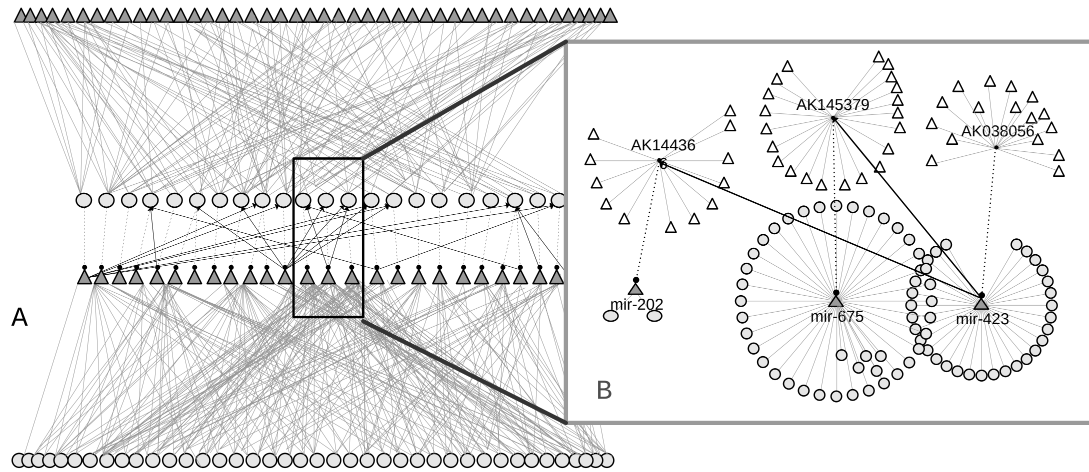


图15. 微型核糖核酸调控信使核糖核酸样核糖核酸网络示意图

参考文献

[1] Fleischmann, R.D., et al., Whole-genome random sequencing and assembly ofHaemophilus influenzae Rd. Science, 1995. 269(5223): p. 496-512.

[2] Finishing the euchromatic sequence of the human genome. Nature, 2004. 431(7011):p.931-45.

[3] Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001.409(6822): p.860-921.

[4] Strogatz, S.H., Exploring complex networks. Nature, 2001. 410(6825): p. 268-76.

[5] Alon, U., Biological networks: the tinkerer as an engineer. Science, 2003. 301(5641):p.1866-7.

[6] http:// www.systemsbiology.org/ Intro_to_ISB_and_Systems_Biology/ Systems_ Biology_ --_the_21st_Century_Science

[7] Mattick, J. S. & Makunin, I. V. Non-coding RNA. Hum Mol Genet 15 Spec No 1, R17-29 (2006).

[8] Liu, C. et al. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res 33, D112-5 (2005).

[9] Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. Nature 450, 304-8 (2007).

[10] Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281-97 (2004).

[11] Lutcke, H. Signal recognition particle (SRP), a ubiquitous initiator of protein translocation. Eur J Biochem 228, 531-50 (1995).

[12] Kiss, T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. Embo J 20, 3617-22 (2001).

[13] Nemes, J. P., Benzow, K. A., Moseley, M. L., Ranum, L. P. & Koob, M. D. The SCA8 transcript is an antisense RNA to a brain-specific transcript encoding a novel actin-binding protein (KLHL1). Hum Mol Genet 9, 1543-51 (2000).

[14] Smilinich, N. J. et al. A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. Proc Natl Acad Sci U S A 96, 8064-9 (1999).

[15] Petrovics, G. et al. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. Oncogene 23, 605-11 (2004).

[16] Ji, P. et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22, 8031-41 (2003).

[17] Numata, K. et al. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res 13, 1301-6 (2003).

[18] Marker, C. et al. Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant Arabidopsis thaliana. Curr Biol 12, 2002-13 (2002).

[19] Huttenhofer, A. et al. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. Embo J 20, 2943-53 (2001).

[20] Rubin, G. M. The draft sequences. Comparing species. Nature 409, 820-1 (2001).

[21] Nadeau, J. H. et al. Sequence interpretation. Functional annotation of mouse genome sequences. Science 291, 1251-5 (2001).

[22] Okazaki, Y. et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563-73 (2002).

[23] Cheng, J. et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149-54 (2005).

[24] Deng, W. et al. Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res 16, 20-9 (2006).

[25] Chen, C. L. et al. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from Oryza sativa. Nucleic Acids Res 31, 2601-13 (2003).

[26] Yuan, G., Klambt, C., Bachellerie, J. P., Brosius, J. & Huttenhofer, A. RNomics in Drosophila melanogaster: identification of 66 candidates for novel non-messenger RNAs. Nucleic Acids Res 31, 2495-507 (2003).

[27] Tang, T. H. et al. Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci U S A 99, 7536-41 (2002).

[28] Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol 11, 1369-73 (2001).

[29] Will, C. L. & Luhrmann, R. Spliceosomal UsnRNP biogenesis, structure and function. Curr Opin Cell Biol 13, 290-301 (2001).

[30] Hinz, S. & Goringer, H. U. The guide RNA database (3.0). Nucleic Acids Res 27, 168 (1999).

[31] Podlevsky, J. D., Bley, C. J., Omana, R. V., Qi, X. & Chen, J. J. The telomerase database. Nucleic Acids Res 36, D339-43 (2008).

[32] Zwieb, C., Gorodkin, J., Knudsen, B., Burks, J. & Wower, J. tmRDB (tmRNA database). Nucleic Acids Res 31, 446-7 (2003).

[33] Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 36, 233-78 (2002).

[34] Rosenblad,M.A., Gorodkin,J., Knudsen,B., Zwieb,C. and Samuelsson,T. (2003) SRPDB:Signal Recognition Particle Database. Nucleic Acids Res., 31, 363-364.

[35] Zwieb,C., Gorodkin,J., Knudsen,B., Burks,J. and Wower,J. (2003) tmRDB (tmRNAdatabase). Nucleic Acids Res., 31, 446-447.

[36] Brown,J.W. (1999) The Ribonuclease P Database. Nucleic Acids Res., 27, 314.

[37] Gu,J., Chen,Y. and Reddy,R.(1998) Small RNA database. Nucleic Acids Res., 26,160-162.

[38] Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2003) Noncoding regulatory RNAsdatabase. Nucleic Acids Res., 31, 429-431.

[39] Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: anRNA family database. Nucleic Acids Res., 31, 439-441.

[40] Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004)GenBank: update. Nucleic Acids Res., 32,D23-D26.

[41] Changning Liu, Baoyan Bai, et al. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res., 33, D112-D115

[42] Kiss,T. (2001) Small nucleolar RNA-guided post-transcriptional modification of cellularRNAs. EMBO J., 20, 3617-3622.

[43] Hendrix,R.W. (1998) Bacteriophage DNA packaging: RNA gears in a DNA transportmachine. Cell, 94, 147-150.

[44] Sugisaki,H. and Takanami,M. (1993) The 5' terminal region of the apocytochrome btranscript in Crithidia fasciculata is successively edited by two guide RNAs in the 3'to 5' direction. J. Biol. Chem., 268, 887-891.

[45] Zhanybekova,S.S.h., Polimbetova,N.S., Nakisbekov,N.O. and Iskakov,B.K. (1996)Detection of a new small RNA, induced by heat shock, in wheat seed ribosomes.Biokhimiia, 61, 862-870.

[46] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) Thetranscriptional landscape of the mammalian genome. Science 309: 1559-1563.

[47] Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencingand characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40-45.

[48] Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, et al. (2006) Transcript annotation inFANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet 2: e62.

[49] Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Globalidentification of human transcribed sequences with genome tiling arrays. Science306: 2242-2246.

[50] Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, et al. (2006)Characterization of the piRNA complex from rat testes. Science 313: 363-367.

[51] Erdmann VA, Szymanski M, Hochberg A, de Groot N, Barciszewski J (1999) Collectionof mRNA-like non-coding RNAs. Nucleic Acids Res 27: 192-195.

[52] Okazaki, Y., et al., Analysis of the mouse transcriptome based on functional annotation of60,770 full-length cDNAs. Nature, 2002. 420(6915): p.563-73.

[53] Ota, T., et al., Complete sequencing and characterization of 21,243 full-length humancDNAs. Nat Genet, 2004. 36(1): p.40-5.

[54] Marahrens Y, Loring J, Jaenisch R (1998) Role of the Xist gene in X chromosomechoosing. Cell 92: 657-664.

[55] Young TL, Matsuda T, Cepko CL (2005) The noncoding RNA taurine upregulated gene 1is required for differentiation of the murine retina. Curr Biol 15: 501-512.

[56] Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, et al. (2005) A strategy forprobing the function of noncoding RNAs finds a repressor of NFAT. Science 309:1570-1573.

[57] Bartel, D.P., MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 2004.116(2): p.281-97.

[58] Ambros, V., The functions of animal microRNAs. Nature, 2004. 431(7006): p.350-5.

[59] Strogatz, S. H. Exploring complex networks. Nature 410, 268-76 (2001).

[60] Bertone, P., et al., Global identification of human transcribed sequences with genome tilingarrays. Science, 2004. 306(5705): p.2242-6.

[61] Ito, T., et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome.Proc Natl Acad Sci U S A, 2001. 98(8): p.4569-74.

[62] Lee, T.I., et al., Transcriptional regulatory networks in Saccharomyces cerevisiae. Science,2002. 298(5594): p.799-804.

[63] Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature 403(6770): 623-7.

[64] Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A 98(8): 4569-74.

[65] Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature 415(6868): 141-7.

[66] Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature 415(6868): 180-3.

[67] Cho, R. J., M. J. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell 2(1): 65-73.

[68] Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression

profiles." Cell 102(1): 109-26.

[69] Tong, A. H., M. Evangelista, et al. (2001). "Systematic genetic analysis with ordered arrays of yeast deletion mutants." Science 294(5550): 2364-8.

[70] Mewes, H. W., D. Frishman, et al. (2002). "MIPS: a database for genomes and protein sequences." Nucleic Acids Res 30(1): 31-4.

[71] Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature 402(6757): 86-90.

[72] Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science 285(5428): 751-3.

[73] Dandekar, T., B. Snel, et al. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem Sci 23(9): 324-8.

[74] Gerdes, S. Y., M. D. Scholle, et al. (2003). "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655." J Bacteriol 185(19): 5673-84.

[75] Huynen, M. A., B. Snel, et al. (2003). "Function prediction and protein networks." Curr Opin Cell Biol 15(2): 191-8.

[76] Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol 18(12): 1257-61.

[77] Hishigaki, H., K. Nakai, et al. (2001). "Assessment of prediction accuracy of protein function from protein--protein interaction data." Yeast 18(6): 523-31.

[78] Maslov, S. and K. Sneppen (2002). "Specificity and stability in topology of protein networks." Science 296(5569): 910-3.

[79] Ge, H., Z. Liu, et al. (2001). "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae." Nat Genet 29(4): 482-6.

[80] Gibson, D., J. Kleinberg, et al. (1998). "Inferring Web communities from link topology." Preceedings of the 9th ACM Conference on Hypertext and Hypermedia ACM Press. New York. NY.

[81] Kleinberg, J. (1998). "Authoritative sources in a hyper-linked environment." Preceedings of the 9th ACM Conference on Hypertext and Hypermedia ACM Press, New York, NY.

[82] von Mering, C., R. Krause, et al. (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature 417(6887): 399-403.

[83] Shen-Orr SS, Milo R, Mangan S, Alon U. (2002) Network motifs in the transcriptionalregulation network of Escherichia coli. Nat Genet 31: 64-68.

[84] Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptionalregulatory networks in Saccharomyces cerevisiae. Science 298: 799-804.

[85] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chclovskii D, et al. (2002) Network motifs:Simple building blocks of complex networks. Science 298: 824-827.

[86] Vazquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, et al. (2004) The topologicalrelationship between the large-scale attributes and local interaction patterns ofcomplex networks. Proc Natl Acad Sci U S A 101: 17940-17945.

[87] Guelzim N, Bottani S, Bourgnie P, Kepes F (2002) Topological and causal structure of theyeast transcriptional regulatory network. Nat Genet 31: 60-63.

[88] Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis ofessentiality within protein networks. Trends Genet 20: 227-231.

[89] Bray D (1995) Protein molecules as computational elements in living cells. Nature 376:307-312.

[90] Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, et al. (2005) Formation ofregulatory patterns during signal propagation in a Mammalian cellular network.Science 309: 1078-1083.

[91] Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, et al. (2004) Genomicanalysis of regulatory network dynamics reveals large topological changes. Nature431: 308-312.

[92] Thomas D, Jacquemin I, Surdin-Kerjan Y (1992) MET4, a leucine zipper protein, andcentromere-binding factor 1 are both required for transcriptional activation of sulfurmetabolism in Saccharomyces cerevisiae. Mol Cell Biol 12: 1719-1727.

[93] Wolf DM, Arkin AP (2003) Motifs, modules, and games in bacteria. Curr Opin Microbiol6: 125.

[94] Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 2005.

[95] Willingham, A. T. et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 309, 1570-3 (2005).

[96] Sood, P., et al. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci U S A 103, 2746-2751

作者简介：

刘长宁：　中国科学院计算技术研究所前瞻研究实验室

孙世伟：　中国科学院计算技术研究所前瞻研究实验室

赵　屹：　中国科学院计算技术研究所前瞻研究实验室，biozy@ict.ac.cn

卜东波：　中国科学院计算技术研究所前瞻研究实验室